HOW DOES UNLABELED DATA IMPROVE Generalization in Self-training? A One-hidden-layer Theoretical Analysis

Shuai Zhang
Rensselaer Polytechnic Institute
Troy, NY, USA 12180
zhangs29@rpi.edu

Meng Wang
Rensselaer Polytechnic Institute
Troy, NY, USA 12180
wangm7@rpi.edu

Sijia Liu
Michigan State University
East Lansing, MI, USA 48824
Sijia Liu, MIT-IBM Watson AI Lab, IBM Research
liusij5@msu.edu

Pin-Yu Chen
IBM Research
Yorktown Heights, NY, USA 10562
Pin-Yu.Chen@ibm.com

Jinjun Xiong
University at Buffalo
Buffalo NY, USA 14260
jinjun@buffalo.edu

ABSTRACT

Self-training, a semi-supervised learning algorithm, leverages a large amount of unlabeled data to improve learning when the labeled data are limited. Despite empirical successes, its theoretical characterization remains elusive. To the best of our knowledge, this work establishes the first theoretical analysis for the known iterative self-training paradigm and proves the benefits of unlabeled data in both training convergence and generalization ability. To make our theoretical analysis feasible, we focus on the case of one-hidden-layer neural networks. However, theoretical understanding of iterative self-training is non-trivial even for a shallow neural network. One of the key challenges is that existing neural network landscape analysis built upon supervised learning no longer holds in the (semi-supervised) self-training paradigm. We address this challenge and prove that iterative self-training converges linearly with both convergence rate and generalization accuracy improved in the order of $1/\sqrt{M}$, where $M$ is the number of unlabeled samples. Experiments from shallow neural networks to deep neural networks are also provided to justify the correctness of our established theoretical insights on self-training.

1 INTRODUCTION

Self-training (Scudder, 1965; Yarowsky, 1995; Lee et al., 2013; Han et al., 2019), one of the most powerful semi-supervised learning (SemiSL) algorithms, augments a limited number of labeled data with unlabeled data so as to achieve improved generalization performance on test data, compared with the model trained by supervised learning using the labeled data only. Self-training has shown empirical success in diversified applications such as few-shot image classification (Su et al., 2020; Xie et al., 2020), object detection (Rosenberg et al., 2005), robustness-aware model training against adversarial attacks (Carmon et al., 2019), continual lifelong learning (Lee et al., 2019), and natural language processing (He et al., 2019; Kahn et al., 2020). The terminology “self-training” has been used to describe various SemiSL
algorithms in the literature, while this paper is centered on the commonly-used iterative self-training method in particular. In this setup, an initial teacher model (learned from the labeled data) is applied to the unlabeled data to generate pseudo labels. One then trains a student model by minimizing the weighted empirical risk of both the labeled and unlabeled data. The student model is then used as the new teacher to update the pseudo labels of the unlabeled data. This process is repeated multiple times to improve the eventual student model. We refer readers to Section 2 for algorithmic details.

Despite the empirical achievement of self-training methods with neural networks, the theoretical justification of such success is very limited, even in the field of SemiSL. The majority of the theoretical results on general SemiSL are limited to linear networks (Chen et al., 2020b; Raghunathan et al., 2020; Oymak & Gulcu, 2020; Oneto et al., 2011). The authors in (Balcan & Blum, 2010) show that unlabeled data can improve the generalization bound if the unlabeled data distribution and target model are compatible. For instance, the unlabeled data need to be well-chosen such that the target function for labeled data can separate the unlabeled data clusters, which, however, may not be able to be verified ahead. Moreover, (Rigollet, 2007; Singh et al., 2008) proves that unlabeled data can improve the convergence rate and generalization error under a similar clustering assumption, where the data contains clusters that have homogeneous labels. A recent work by Wei et al. (2020) analyzes SemiSL on nonlinear neural networks and proves that an infinite number of unlabeled data can improve the generalization compared with training with labeled data only. However, Wei et al. (2020) considers single shot rather than iterative SemiSL, and the training problem aims to minimize the consistency regularization rather than the risk function in the conventional self-training method (Lee et al., 2013). Moreover, Wei et al. (2020) directly analyzes the global optimum of the nonconvex training problem without any discussion about how to achieve the global optimum. To the best of our knowledge, there exists no analytical characterization of how the unlabeled data affect the generalization of the learned model by iterative self-training on nonlinear neural networks.

**Contributions.** This paper provides the first theoretical study of iterative self-training on nonlinear neural networks. Focusing on one-hidden-layer neural networks, this paper provides a quantitative analysis of the generalization performance of iterative self-training as a function of the number of labeled and unlabeled samples. Specifically, our contributions include

1. **Quantitative justification of generalization improvement by unlabeled data.** Assuming the existence of a ground-truth model with weights $W^*$ that maps the features to the corresponding labels, we prove that the learned model via iterative self-training moves closer to $W^*$ as the number $M$ of unlabeled data increases, indicating a better testing performance. Specifically, we prove that the Frobenius distance to $W^*$, which is approximately linear in the generalization error, decreases in the order of $1/\sqrt{M}$. As an example, Figure 1 shows that the proposed theoretical bound matches the empirical self-training performance versus the number of unlabeled data for image classification; see details in Section 4.2.

2. **Analytical justification of iterative self-training over single shot alternative.** We prove that the student models returned by the iterative self-training method converge linearly to a model close to $W^*$, with the rate improvement in the order of $1/\sqrt{M}$.

3. **Sample complexity analysis of labeled and unlabeled data for learning a proper model.** We quantify the impact of labeled and unlabeled data on the generalization of the learned model. In particular, we prove that the sample complexity of labeled data can be reduced compared with supervised learning.

**1.1 RELATED WORKS**

**Semi-supervised learning.** Besides self-training, many recent SemiSL algorithms exploit either consistency regularization or entropy minimization. Consistency regularization is based on the as-
assumption that the learned model will return same or similar output when the input is perturbed (Laine & Aila, 2016; Bachman et al., 2014; Sajjadi et al., 2016; Tarvainen & Valpola, 2017; Reed et al., 2015). (Grandvalet & Bengio, 2005) claims that the unlabeled data are more informative if the pseudo labels of the unlabeled data have lower entropy. Therefore, a line of works (Grandvalet & Bengio, 2005; Miyato et al., 2018) adds a regularization term that minimizes the entropy of the outputs of the unlabeled data. In addition, hybrid algorithms that unify both the above regularizations have been developed like (Berthelot et al., 2019a; Sohn et al., 2020).

**Domain adaptation.** Domain adaptation exploits abundant data in the source domain to learn a model for the target domain, where only limited training data are available (Liebelt & Schmid, 2010; Vazquez et al., 2013; Zhang et al., 2013; Long et al., 2015; Tzeng et al., 2014). Source and target domain are related but different. Unsupervised domain adaptation (Ganin & Lempitsky, 2015; Ganin et al., 2016; Gong et al., 2013; Bousmalis et al., 2016), where training data in target domain are unlabeled, is similar to SemiSL, and self-training methods have been used for analysis (Zou et al., 2018; Tang et al., 2012; French et al., 2019; Zhang et al., 2016; Brutzkus & Globerson, 2017). However, self-training and unsupervised domain adaptation are fundamentally different. The former learns a model for the domain where there is limited labeled data, with the help of a large number of unlabeled data from a different domain. The latter learns a model for the domain where the training data are unlabeled, with the help of sufficient labeled data from a different domain.

**Generalization analysis of supervised learning.** In theory, the testing error is upper bounded by the training error plus the generalization gap between training and testing. These two quantities are often analyzed separately and cannot be proved to be small simultaneously for deep neural networks. For example, neural tangent kernel (NTK) method (Jacot et al., 2018; Du et al., 2018; Lee et al., 2018) shows the training error can be zero, and the Rademacher complexity in (Bartlett & Mendelson, 2002) bounds the generalization gap (Arora et al., 2019a). For one-hidden-layer neural networks (Safran & Shamir, 2018), the testing error can be proved to be zero under mild conditions. One common assumption is that the input data belongs to the Gaussian distribution (Zhong et al., 2017; Ge et al., 2018; Kalai et al., 2008; Bakshi et al., 2019; Zhang et al., 2016; Brutzkus & Globerson, 2017; Li & Yuan, 2017; Soltanolkotabi et al., 2018). Another line of approaches (Brutzkus et al., 2018; Li & Liang, 2018; Wang et al., 2019) consider linearly separable data.

The rest of this paper is organized as follows. Section 2 introduces the problem formulation and self-training algorithm. Major results are summarized in Section 3 and empirical evaluations are presented in Section 4. Section 5 concludes the whole paper. All the proofs are in the Appendix.

### 2 Formalizing Self-Training: Notation, Formulation, and Algorithm

**Problem formulation.** Given $N$ labeled data sampled from distribution $P_{l}$, denoted by $\mathcal{D} = \{x_n, y_n\}_{n=1}^N$, and $M$ unlabeled data drawn from distribution $P_{u}$, denoted by $\mathcal{D} = \{\tilde{x}_m\}_{m=1}^M$. The aim is to find a neural network model $g(W)$, where $W$ denotes the trainable weights, that minimizes the testing error on data sampled from $P_{l}$.

**Table 1: Iterative Self-Training**

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>Initialize iteration $\ell = 0$ and obtain a model $W^{(\ell)}$ as the teacher using labeled data $\mathcal{D}$ only;</td>
</tr>
<tr>
<td>S2</td>
<td>Use the teacher model to obtain pseudo labels $\tilde{y}_m$ of unlabeled data in $\mathcal{D}$;</td>
</tr>
<tr>
<td>S3</td>
<td>Train the neural network by minimizing (1) via $T$-step mini-batch gradient descent method using disjoint subsets ${D_t}_{t=0}^{T-1}$ and ${\tilde{D}<em>t}</em>{t=0}^{T-1}$ of $\mathcal{D}$. Let $W^{(\ell+1)}$ denote the obtained student model;</td>
</tr>
<tr>
<td>S4</td>
<td>Use $W^{(\ell+1)}$ as the current teacher model. Let $\ell \leftarrow \ell + 1$ and go back to step S2;</td>
</tr>
</tbody>
</table>

**Iterative self-training.** In each iteration, given the current teacher predictor $g(W^{(\ell)})$, the pseudo-labels for the unlabeled data in $\mathcal{D}$ are computed as $\tilde{y}_m = g(W^{(\ell)}; \tilde{x}_m)$. The method then minimizes the weighted empirical risk $\tilde{f}_{\mathcal{D}, \tilde{\mathcal{D}}}(W)$ of both labeled and unlabeled data through stochastic gradient
descent, where
\[
\hat{J}_D, \hat{J}(\mathbf{W}) = \frac{\lambda}{2N} \sum_{n=1}^{N} (y_n - g(\mathbf{W}; \mathbf{x}_n))^2 + \frac{\bar{\lambda}}{2M} \sum_{m=1}^{M} (\bar{y}_m - g(\mathbf{W}; \bar{\mathbf{x}}_m))^2,
\]
and \(\lambda + \bar{\lambda} = 1\). The learned student model \(g(\mathbf{W}^{(t+1)})\) is used as the teacher model in the next iteration. The initial model \(g(\mathbf{W}^{(0)})\) is learned from labeled data. The formal algorithm is summarized as in Table 1.

**Model and assumptions.** This paper considers regression where \(g\) is a one-hidden-layer fully connected neural network equipped with \(K\) neurons. Namely, given the input \(\mathbf{x} \in \mathbb{R}^d\) and weights \(\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K] \in \mathbb{R}^{d \times K}\), we have
\[
g(\mathbf{W}; \mathbf{x}) := \frac{1}{K} \sum_{j=1}^{K} \phi(w_j^T \mathbf{x}),
\]
where \(\phi\) is the ReLU activation function and \(\phi(z) = \max\{z, 0\}\) for any input \(z \in \mathbb{R}\). Here, we fix the top layer weights as 1 for simplicity, and the equivalence of such a simplification is discussed in Appendix K.

Moreover, we assume an unknown ground-truth model with weights \(\mathbf{W}^*\) that maps all the features to the corresponding labels drawn from \(P_t\), i.e., \(y = g(\mathbf{W}^*; \mathbf{x})\), where \((\mathbf{x}, y) \sim P_t\). The generalization function (GF) with respect to \(g(\mathbf{W})\) is defined as
\[
I(\mathbf{g}(\mathbf{W})) = \mathbb{E}_{(\mathbf{x}, y) \sim P_t} (y - g(\mathbf{W}; \mathbf{x}))^2 = \mathbb{E}_{(\mathbf{x}, y) \sim P_t} (g(\mathbf{W}^*; \mathbf{x}) - g(\mathbf{W}; \mathbf{x}))^2.
\]
By definition \(I(\mathbf{g}(\mathbf{W}^*))\) is zero. Clearly, \(\mathbf{W}^*\) is not unique because any column permutation of \(\mathbf{W}^*\), which corresponds to permuting neurons, represents the same function as \(\mathbf{W}^*\) and minimizes GF in (3). To simplify the representation, we follow the convention and abuse the notation that the distance from \(\mathbf{W}\) to \(\mathbf{W}^*\), denoted by \(\|\mathbf{W} - \mathbf{W}^*\|_F\), means the smallest distance from \(\mathbf{W}\) to any permutation of \(\mathbf{W}^*\). Additionally, some important notations are summarized in Table 2.

We assume the inputs of both the labeled and unlabeled data belong to the zero mean Gaussian distribution, i.e., \(\mathbf{x} \sim \mathcal{N}(0, \delta^2 \mathbf{I}_d)\), and \(\bar{\mathbf{x}} \sim \mathcal{N}(0, \delta^2 \mathbf{I}_d)\). The Gaussian assumption is motivated by the data whitening [LeCun et al., 2012] and batch normalization techniques [Ioffe & Szegedy, 2015] that are commonly used in practice to improve learning performance. Moreover, training one-hidden-layer neural network with multiple neurons is NP-Complete [Blum & Rivest, 1992] without any assumption.

**The focus of this paper.** This paper will analyze three aspects about self-training: (1) the generalization performance of \(\mathbf{W}^{(L)}\), the returned model by self-training after \(L\) iterations, measured by \(\|\mathbf{W}^{(L)} - \mathbf{W}^*\|_F\); (2) the influence of parameter \(\lambda\) in (1) on the training performance; and (3) the impact of unlabeled data on the training and generalization performance.

<table>
<thead>
<tr>
<th>Table 2: Some Important Notations</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\mathcal{D} = {\mathbf{x}<em>n, y_n}</em>{n=1}^{N})</td>
</tr>
<tr>
<td>(\mathcal{D} = {\tilde{\mathbf{x}}<em>m}</em>{m=1}^{M})</td>
</tr>
<tr>
<td>(d)</td>
</tr>
<tr>
<td>(K)</td>
</tr>
<tr>
<td>(\kappa)</td>
</tr>
<tr>
<td>(\mathbf{W}^{(t)})</td>
</tr>
<tr>
<td>(\mathbf{W}^*)</td>
</tr>
<tr>
<td>(\mathbf{W}^{[\lambda]})</td>
</tr>
</tbody>
</table>

\(^1\)The results can be extended to binary classification with a cross-entropy loss function. Please see Appendix I.

\(^2\)Because ReLU is non-linear and non-smooth, (1) is non-convex and non-smooth, which poses analytical challenges. The results can be easily extended to smooth functions with bounded gradients, e.g., Sigmoid.

\(^3\)We use this metric because \(I(\mathbf{g}(\mathbf{W}))\) is shown to be linear in \(\|\mathbf{W}^{(L)} - \mathbf{W}^*\|_F\) numerically when \(\mathbf{W}^{(L)}\) is close to \(\mathbf{W}^*\), see Figure 4.
Beyond supervised learning: Challenge of self-training. The existing theoretical works such as (Zhong et al., 2017) [Zhang et al., 2020a,b,c] verify that for one-hidden-layer neural networks, if only labeled data are available, and $x$ are drawn from the standard Gaussian distribution, then supervised learning by minimizing (1) with $\lambda = 1$ can return a model with ground-truth weights $W^*$ (up to column permutation), as long as the number of labeled data $N$ is at least $N^*$, which depends on $\kappa, K$ and $d$. In contrast, this paper focuses on the low labeled-data regime when $N$ is less than $N^*$. Specifically,

$$N^*/4 < N \leq N^*. \quad (4)$$

Intuitively, if $N < N^*$, the landscape of the empirical risk of the labeled data becomes highly non-convex, even in a neighborhood of $W^*$, thus, the existing analyses for supervised learning do not hold in this region. With additional unlabeled data, the landscape of the weighted empirical risk becomes smoother near $W^*$. Moreover, as $M$ increases, and starting from a nearby initialization, the returned model $W^{(L)}$ by iterative self-training can converge to a local minimum that is closer to $W^*$ (see illustration in Figure 2).

Compared with supervised learning, the formal analyses of self-training need to handle new technical challenges from two aspects. First, the existing analyses of supervised learning exploit the fact that the GF and the empirical risk have the same minimizer, i.e., $W^*$. This property does not hold for self-training as $W^*$ no longer minimizes the weighted empirical risk in (1). Second, the iterative manner of self-training complicates the analyses. Specifically, the empirical risk in each iteration is different and depends on the model trained in the previous iteration through the pseudo labels.

In what follows, we provide theoretical insights and the formal theorems. Some important quantities $\lambda$ and $\mu$ are defined below

$$\hat{\lambda} := \frac{\lambda \delta^2}{\lambda \delta^2 + \delta \lambda^2}, \quad \text{and} \quad \mu(\delta, \hat{\delta}) := \sqrt{\frac{\lambda \delta^2 + \lambda \hat{\delta}^2}{\lambda \rho(\delta) + \lambda \rho(\hat{\delta})}}, \quad (5)$$

where $\rho$ is a positive function defined in (73). $\hat{\lambda}$ is an increasing function of $\lambda$. Also, from Lemma 1 (in Appendix), $\rho(\delta)$ is in the order of $\delta^2$ when $\delta \leq 1$ for ReLU activation functions. Thus, $\mu$ is a fixed constant, denoted by $\mu^*$, for all $\delta, \hat{\delta} < 1$. When $\delta$ and $\hat{\delta}$ are large, $\mu$ increases as they increase.

The formal definition of $N^*$ in (4) is $c(\kappa)\mu^2 K^3 d \log q$, where $c(\kappa)$ is some polynomial function of $\kappa$ and can be viewed as constant.

3.1 Informal key theoretical findings

To the best of our knowledge, Theorems 1 and 2 provide the first theoretical characterization of iterative self-training on nonlinear neural networks. Before formally presenting them, we summarize the highlights as follows.

1. Linear convergence of the learned models. The learned models converge linearly to a model close to $W^*$. Thus, the iterative approach returns a model with better generalization than that by the single-shot method. Moreover, the convergence rate is a constant term plus a term in the order of $1/\sqrt{M}$ (see $\Delta_1$ in Figure 3), indicating a faster convergence with more unlabeled data.

$$\Delta_1 = \frac{1 + O\left(\frac{1}{\sqrt{M}}\right)}{M} \frac{\lambda^2}{\kappa}$$

$$\Delta_1 = \frac{1 + O\left(\frac{1}{\sqrt{M}}\right)}{M} \frac{\lambda^2}{\kappa}$$

$$\Delta_4 = \frac{1 + O\left(\frac{1}{\sqrt{M}}\right)}{M} \frac{\lambda^2}{\kappa}$$

$$\Delta_4 = \frac{1 + O\left(\frac{1}{\sqrt{M}}\right)}{M} \frac{\lambda^2}{\kappa}$$

Figure 3: Illustration of the (1) ground truth $W^*$, (2) iterations $\{W^{(i)}\} \in \mathbb{R}^n$ for $i = 0$, (3) convergent point $W^{(L)}$, and (4) $W[\lambda] = \hat{\lambda}W^* + (1 - \hat{\lambda})W^{(0)}$. 

Figure 2: Adding unlabeled data in the empirical risk function drives its local minimum closer to $W^*$, which minimizes the generalization function.
2. Returning a model with guaranteed generalization in the low labeled-data regime. Even when the number of labeled data is much less than the required sample complexity to obtain \(W^*\) in supervised learning, we prove that with the help of unlabeled data, the iterative self-training can return a model in the neighborhood of \(W^{[\lambda]}\), where \(W^{[\lambda]}\) is in the line segment of \(W^{(0)}(\hat{\lambda} = 0)\) and ground truth \(W^*(\hat{\lambda} = 1)\). Moreover, \(\hat{\lambda}\) is upper bounded by \(\sqrt{N/N^*}\). Thus \(W^{(L)}\) moves closer to \(W^*\) as \(N\) increases (\(\mathcal{E}_0\) in Figure 3), indicating a better generalization performance with more labeled data.

3. Guaranteed generalization improvement by unlabeled data. The distance between \(W^{(L)}\) and \(W^{[\lambda]}\) (\(\mathcal{E}_1\) in Figure 3) scales in the order of \(1/\sqrt{M}\). With a larger number of unlabeled data \(M\), \(W^{(L)}\) moves closer to \(W^{[\lambda]}\) and thus \(W^*\), indicating an improved generalization performance (Theorem 1). When \(N\) is close to \(N^*\) but still smaller as defined in (12), both \(W^{(L)}\) and \(W^{[\lambda]}\) converge to \(W^*\), and thus the learned model achieves zero generalization error (Theorem 2).

3.2 Formal Theory in Low Labeled-Data Regime

Takeaways of Theorem 1: Theorem 1 characterizes the convergence rate of the proposed algorithm and the accuracy of the learned model \(W^{(L)}\) in a low labeled-data regime. Specifically, the iterates converge linearly, and the learned model is close to \(W^{[\lambda]}\) and guaranteed to outperform the initial model \(W^{(0)}\).

**Theorem 1.** Suppose the initialization \(W^{(0)}\) and the number of labeled data satisfy

\[
\|W^{(0)} - W^*\|_F \leq p^{-1} \cdot \frac{\|W^*\|_F}{c(\kappa)\mu^2 K^{3/2}} \quad \text{with} \quad p \in \left(\frac{1}{2}, 1\right],
\]

and

\[
\max\left\{\frac{1}{K}, p - \frac{2p - 1}{\mu \sqrt{K}}\right\} \leq \lambda \leq \min\left\{\sqrt{\frac{N}{N^*}}, p + \frac{2p - 1}{\mu \sqrt{K}}\right\},
\]

If the value of \(\hat{\lambda}\) in (5) and unlabeled data amount \(M\) satisfy

\[
\max\left\{\frac{1}{K}, p - \frac{2p - 1}{\mu \sqrt{K}}\right\} \leq \hat{\lambda} \leq \min\left\{\sqrt{\frac{N}{N^*}}, p + \frac{2p - 1}{\mu \sqrt{K}}\right\},
\]

and

\[
M \geq (2p - 1)^{-2} c(\kappa) \mu^2 (1 - \hat{\lambda})^2 K^3 d \log q.
\]

Then, when the number \(T\) of SGD iterations is large enough in each loop \(\ell\), with probability at least \(1 - q^{-d}\), the iterates \(\{W^{(\ell)}\}_{\ell=0}^L\) converge to \(W^{[\lambda]}\) as

\[
\|W^{(L)} - W^{[\lambda]}\|_F \leq \left(1 + \Theta\left(\frac{\mu(1 - \hat{\lambda})}{\sqrt{M}}\right)\right) \cdot \|W^{(0)} - W^{[\lambda]}\|_2 + \left(1 + \Theta\left(\frac{\mu(1 - \hat{\lambda})}{\sqrt{M}}\right)\right) \cdot \|W^* - W^{[\lambda]}\|_F,
\]

where \(W^{[\lambda]} = \hat{\lambda}W^* + (1 - \hat{\lambda})W^{(0)}\). Typically, when the iteration number \(L\) is sufficient large, we have

\[
\|W^{(L)} - W^*\|_F \leq \left(1 + \Theta\left(\frac{\mu(1 - \hat{\lambda})}{\sqrt{M}}\right)\right) \cdot 2(1 - \hat{\lambda}) \cdot \|W^* - W^{(0)}\|_F.
\]

The accuracy of the learned model \(W^{(L)}\) with respect to \(W^*\) is characterized as (11), and the learning model is better than initial model as in (11) if the following conditions hold. First, the weights \(\lambda\) in (8) are properly chosen as in (8). Second, the number of unlabeled data is sufficiently large as in (9).

**Selection of \(\lambda\) in self-training algorithms.** When \(\hat{\lambda}\) increases, the required number of unlabeled data is reduced from (9), and the convergence point \(W^{(L)}\) becomes closer to \(W^*\) from (11), which indicates a smaller generalization error. Thus, a large \(\hat{\lambda}\) within its feasible range (9) is desirable. When the initial model \(W^{(0)}\) is closer to \(W^*\) (corresponding to a larger \(p\)), and the number of labeled data \(N\) increases, the upper bound in (8) increases, and thus, one can select a larger \(\hat{\lambda}\).

**The initial model \(W^{(0)}\)** The tensor initialization from (Zhong et al., 2017) can return \(W^{(0)}\) that satisfies (6) when the number of labeled data is \(N = p^2 N^*\) (see Lemma 3 in Appendix). Combining with the requirement in (7), Theorem 1 applies to the case that \(N\) is at least \(N^*/4\).
3.3 Formal theory of achieving zero generalization error

**Takeaways of Theorem 2:** Theorem 2 indicates the model returned by the proposed algorithm converges linearly to the ground truth $W^*$. Thus the distance between the learned model and the ground truth can be arbitrarily small with the ability to achieve zero generalization error. The required sample complexity is reduced by a constant factor compared with supervised learning.

**Theorem 2.** Consider the number of unlabeled data satisfies

$$(1 - 1/(\mu \sqrt{K}))^2 \cdot N^* \leq N \leq N^*,$$  \hspace{1cm} (12)

we choose $\hat{\lambda}$ such that

$$1 - 1/(\mu \sqrt{K}) \leq \hat{\lambda} \leq \sqrt{N/N^*}.$$  \hspace{1cm} (13)

Suppose the initial model $W^{(0)}$ and the number of unlabeled data $M$ satisfy

$$\|W^{(0)} - W^*\|_F \leq \frac{\|W^*\|_F}{c(\kappa)\mu^2K^{3/2}} \quad \text{and} \quad M \geq c(\kappa)\mu^2(1 - \hat{\lambda})^2K^3d \log q,$$  \hspace{1cm} (14)

the iterates $\{W^{(L)}\}_{L=0}^L$ converge to the ground truth $W^*$ as follows,

$$\|W^{(L)} - W^*\|_F \leq \left(1 + \frac{c(\kappa)\hat{\lambda}}{\sqrt{N}} + \frac{c(\kappa)(1 - \hat{\lambda})}{\sqrt{M}}\right) \cdot \mu \sqrt{K}(1 - \hat{\lambda})^L \cdot \|W^{(0)} - W^*\|_F.$$  \hspace{1cm} (15)

The models $W^{(L)}$'s converge linearly to the ground truth $W^*$ as (15) when the number of labeled data satisfies (12). In contrast, supervised learning requires at least $N^*$ labeled samples to estimate $W^*$ accurately without unlabeled data, which suggests self-training at least saves a constant fraction of labeled data.

3.4 The main proof idea

Our proof builds upon and extends one recent line of works on supervised learning such as (Zhong et al. 2017; Zhang et al. 2020b,c; 2021). The standard framework of these works is first to show that the generalization function $I(g(W))$ in (3) is locally convex near $W^*$, which is its global minimizer. Then, when $M = 0$ and $N$ is sufficiently large, the empirical risk function using labeled data only can approximate $I(g(W))$ well in the neighborhood of $W^*$. Thus, if initialized in this local convex region, the iterations, returned by applying gradient descent approach on the empirical risk function, converge to $W^*$ linearly.

The technical challenge here is that in self-training, when unlabeled data are paired with pseudo labels, $W^*$ is no longer a global minimizer of the empirical risk $\hat{f}_{D,\hat{\lambda}}$ in (1), and $\hat{f}_{D,\hat{\lambda}}$ does not approach $I(g(W))$ even when $M$ and $N$ increase to infinity. Our new idea is to design a population risk function $f(W; \hat{\lambda})$ in (17) (see appendix), which is a lower bound of $\hat{f}_{D,\hat{\lambda}}$ when $M$ and $N$ are infinity. $f(W; \hat{\lambda})$ is locally convex around its minimizer $W^{[\hat{\lambda}]}$, and $W^{[\hat{\lambda}]}$ approaches $W^*$ as $\hat{\lambda}$ increases. Then we show the iterates generated by $\hat{f}_{D,\hat{\lambda}}$ stay close to $f(W; \hat{\lambda})$, and the returned model $W^{(L)}$ is close to $W^{[\hat{\lambda}]}$. New technical tools are developed to bound the distance between the functions $\hat{f}_{D,\hat{\lambda}}$ and $f(W; \hat{\lambda})$.

4 Empirical results

4.1 Synthetic data experiments

We generate a ground-truth neural network with the width $K = 10$. Each entry of $W^*$ is uniformly selected from $[-2.5, 2.5]$. The input of labeled data $x_n$ are generated from Gaussian distribution $N(0, I_d)$ independently, and the corresponding label $y_n$ is generated through (2) using $W^*$. The unlabeled data $\tilde{x}_m$ are generated from $N(0, \delta^2 I_d)$ independently with $\delta = 1$ except in Figure 7. $d$ is set as 50 except in Figure 9. The value of $\lambda$ is selected as $\sqrt{N/(2Kd)}$ except in Figure 8.

We consider one-hidden-layer except in Figure 4. The initial teacher model $W^{(0)}$ in self-training is randomly selected from $\{W||W - W^*||_F \leq 0.5\}$ to reduce the computation. In
each iteration, the maximum number of SGD steps $T$ is $10$. Self-training terminates if $||W^{(t)} - W^{(t+1)}||_F / ||W^{(t)}||_F \leq 10^{-4}$ or reaching 1000 iterations. In Figures 5 to 8, all the points on the curves are averaged over 1000 independent trials, and the regions in lower transparency indicate the corresponding one-standard-deviation error bars. Our empirical observations are summarized below.

(a) GF (testing performance) proportional to $||W - W^*||_F$. Figure 4 illustrates the GF in (3) against the distance to the ground truth $W^*$. To visualize results for different networks together, GF is normalized in $[0, 1]$, divided by its largest value for each network architecture. All the results are averaged over 100 independent choices of $W$. One can see that for one-hidden-layer neural networks, in a large region near $W^*$, GF is almost linear in $||W - W^*||_F$. When the number of hidden layers increases, this region decreases, but the linear dependence still holds locally. This is an empirical justification of using $||W - W^*||_F$ to evaluate the GF and, thus, the testing error in Theorems 1 and 2.

(b) $||W^{(L)} - W^*||_F$ as a linear function of $1/\sqrt{M}$. Figure 5 shows the relative error $||W^{(L)} - W^*||_F / ||W^*||_F$ when the number of unlabeled data and labeled data changes. One can see that the relative error decreases when either $M$ or $N$ increases. Additionally, the dash-dotted lines represent the best fitting of the linear functions of $1/\sqrt{M}$ using the least square method. Therefore, the relative error is indeed a linear function of $1/\sqrt{M}$, as predicted by our results in (11) and (15).

(c) Convergence rate as a linear function of $1/\sqrt{M}$. Figure 6 illustrates the convergence rate when $M$ and $N$ change. We can see that the convergence rate is a linear function of $1/\sqrt{M}$, as predicted by our results (11) and (15). When $M$ increases, the convergence rate is improved, and the method converges faster.

(d) Increase of $\delta$ slows down convergence. Figure 7 shows that the convergence rate becomes worse when the variance of the unlabeled data $\delta$ increases from 1. When $\delta$ is less than 1, the convergence rate almost remains the same, which is consistent with our characterization in (10) that the convergence rate is linear in $\mu$. From the discussion after (5), $\mu$ increases as $\delta$ increases from 1 and stays constant when $\delta$ is less than 1.

(e) $||W^{(L)} - W^*||_F / ||W^*||_F$ is improved as a linear function of $\hat{\lambda}$. Figure 8 shows that the relative errors of $W^{(L)}$ with respect to $W^*$ decrease almost linearly when $\hat{\lambda}$ increases, which is consistent with the theoretical result in (11). Moreover, when $\hat{\lambda}$ exceeds a certain threshold positively correlated with $N$, the relative error increases rather than decreases. That is consistent with the analysis in (8) that $\hat{\lambda}$ has an upper limit, and such a limit increases as $N$ increases.

(f) Unlabeled data reduce the sample complexity to learn $W^*$. Figure 9 depicts the phase transition of returning $W^{(L)}$. For every pair of $d$ and $N$, we construct 100 independent trials, and each trial is said to be successful if $||W^{(L)} - W^*||_F / ||W^*||_F \leq 10^{-2}$. The white blocks correspond to the successful trials, while the block in black indicates all failures. When $d$ increases, the required number of labeled data to learn $W^*$ is linear in $d$. Thus, the sample complexity bound in (12) is order-wise optimal for $d$. Moreover, the phase transition line when $M = 1000$ is below the one when $M = 0$. Therefore, with unlabeled data, the required sample complexity of $N$ is reduced.
Figure 7: Convergence rate with different $\delta$.

Figure 8: $\|W^{(L)} - W^*\|_F/\|W^*\|_F$ when $\hat{\lambda}$ and $N$ change.

Figure 9: Empirical phase transition of the curves with (a) $M = 0$ and (b) $M = 1000$.

4.2 IMAGE CLASSIFICATION ON AUGMENTED CIFAR-10 DATASET

We evaluate self-training on the augmented CIFAR-10 dataset, which has 50K labeled data. The unlabeled data are mined from 80 Million Tiny Images following the setup in (Carmon et al., 2019) and additional 50K images are selected for each class, which is a total of 500K images, to form the unlabeled data. The self-training method is the same implementation as that in (Carmon et al., 2019). $\lambda$ and $\hat{\lambda}$ is selected as $N/(M + N)$ and $M/(N + M)$, respectively, and the algorithm stops after 200 epochs. In Figure 10, the dash lines stand for the best fitting of the linear functions of $1/\sqrt{M}$ via the least square method. One can see that the test accuracy is improved by up to 7% using unlabeled data, and the empirical evaluations match the theoretical predictions. Figure 11 shows the convergence rate calculated based on the first 50 epochs, and the convergence rate is almost a linear function of $1/\sqrt{M}$, as predicted by (10).

Figure 10: The test accuracy against the number of unlabeled data

Figure 11: The convergence rate against the number of unlabeled data

5 CONCLUSION

This paper provides new theoretical insights into understanding the influence of unlabeled data in the iterative self-training algorithm. We show that the improved generalization error and convergence rate is a linear function of $1/\sqrt{M}$, where $M$ is the number of unlabeled data. Moreover, compared with supervised learning, using unlabeled data reduces the required sample complexity of labeled data for achieving zero generalization error. Future directions include generalizing the analysis to multi-layer neural networks and other semi-supervised learning problems such as domain adaptation.

ACKNOWLEDGEMENT

This work was supported by AFOSR FA9550-20-1-0122, ARO W911NF-21-1-0255, NSF 1932196 and the Rensselaer-IBM AI Research Collaboration (http://airc.rpi.edu), part of the IBM AI Horizons Network (http://ibm.biz/AIHorizons).

The codes are downloaded from https://github.com/yaircarmon/semisup-adv
REFERENCES


Appendix

A Overview of the Proof Techniques

We first provide an overview of the techniques used in proving Theorems 1 and 2.

1. Characterization of a proper population risk function. To characterize the performance of the iterative self-training algorithm via the stochastic gradient descent method, we need first to define a population risk function such that the following two properties hold. First, the landscape of the population risk function should be analyzable near \( \{W^{(t)}\}_{t=0} \). Second, the distance between the empirical risk function in (1) and the population risk function should be bounded near \( \{W^{(t)}\}_{t=0} \). The generalization function defined in (3), which is widely used in the supervised learning problem with a sufficient number of samples, failed the second requirement. To this end, we turn to find a new population risk function defined in (17), and the illustrations of the population risk function and objection function are included in Figure 12.

2. Local convex region of the population risk function. The purpose is to characterize the iterations via the stochastic gradient descent method in the population risk function. To obtain the local convex region of the population risk function, we first bound the Hessian of the population risk function at its global optimal. Then, we utilize Lemma 12 in Appendix H.1 to obtain the Hessian of the population risk function near the global optimal. The local convex region of the population risk function is summarized in Lemma 1, and the proof of Lemma 1 is included in Appendix H.1.

3. Bound between the population risk and empirical risk functions. After the characterization of the iterations via the stochastic gradient descent method in the population risk function, we need to bound the distance between the population risk function and empirical risk function. Therefore, the behaviors of the iterations via the stochastic gradient descent method in the empirical risk function can be described by the ones in the population risk function and the distance between these two. The key lemma is summarized in Lemma 2 (see Appendix H.2), and the proof is included in Appendix H.2.

![Figure 12: The landscapes of the objection function and population risk function.](image)

In the following contexts, the details of the iterative self-training algorithm are included in Appendix B. We then first provide the proof of Theorem 2 in Appendix E, which can be viewed as a special case of Theorem 1. Then, with the preliminary knowledge from proving Theorem 2, we turn to present the full proof of a more general statement summarized in Theorem 3 (see Appendix F), which is related to Theorem 1. The definition and relative proofs of \( \mu \) and \( \rho \) are all included in Appendix C. The proofs of preliminary lemmas are included in Appendix H.

B Iterative Self-Training Algorithm

In this section, we implement the details of the mini-batch stochastic gradient descent used in each stage of the iterative self-training algorithm. After \( t \) number of iterations via mini-batch stochastic...
gradient descent at \( \ell \)-th stage of self-training algorithm, the learned model is denoted as \( W^{(\ell,t)} \). One can easily check that \( W^{(t)} \) in the main context is denoted as \( W^{(t,0)} \) in this section and the following proofs. Last, the pseudo-code of the iterative self-training algorithm is summarized in Algorithm 1.

### Algorithm 1 Iterative Self-Training Algorithm

**Input:** labeled \( \mathcal{D} = \{ (\mathbf{x}_n, y_n) \}_{n=1}^N \), unlabeled data \( \mathcal{D} = \{ \mathbf{x}_m \}_{m=1}^M \), and gradient step size \( \eta \);

**Initialization:** preliminary teacher model with weights \( W^{(0,0)} \);

**Partition:** randomly and independently pick data from \( \mathcal{D} \) and \( \mathcal{D} \) to form \( T \) subsets \( \{ D_t \}_{t=0}^{T-1} \) and \( \{ \tilde{D}_t \}_{t=0}^{T-1} \), respectively;

for \( \ell = 0, 1, \cdots, L-1 \) do

\[
y_m = g(W^{(\ell,0)}; \mathbf{x}_m) \quad \text{for} \quad m = 1, 2, \cdots, M
\]

for \( t = 0, 1, \cdots, T-1 \) do

\[
W^{(\ell,t+1)} = W^{(\ell,t)} - \eta \cdot \nabla \hat{f}_{D_t, \tilde{D}_t}(W^{(\ell,t)}) + \beta \cdot (W^{(\ell,t)} - W^{(\ell,t-1)})
\]

end for

\[
W^{(\ell+1,0)} = W^{(\ell,T)}
\]

end for

### C Notations

In this section, we first introduce some important notations that will be used in the following proofs, and the notations are summarized in Table 1.

As shown in Algorithm 1, \( W^{(\ell,t)} \) denotes the learned model after \( t \) number of iterations via mini-batch stochastic gradient descent at \( \ell \)-th stage of the iterative self-training algorithm. Given a student model \( \hat{W} \), the pseudo label for \( \mathbf{x} \in \tilde{D} \) is generated as

\[
\hat{y} = g(\hat{W}; \mathbf{x}).
\]

Further, let \( W^{[p]} = pW^* + (1-p)W^{(0,0)} \), we then define the population risk function as

\[
f(W; p) = \frac{\lambda}{2} E_x \left( y^*(p) - g(W; x) \right)^2 + \frac{\gamma}{2} E_{\bar{x}} \left( \bar{y}^*(p) - g(W; \bar{x}) \right)^2,
\]

where \( y^*(p) = g(W^{[p]}; x) \) with \( x \sim \mathcal{N}(0, \delta^2 I) \) and \( \bar{y}^*(p) = g(W^{[p]}; \bar{x}) \) with \( \bar{x} \sim \mathcal{N}(0, \bar{\delta}^2 I) \). When \( p = 1 \), we have \( W^{[p]} = W^* \) and \( y^*(p) = y \) for data in \( D \).

Moreover, we use \( \sigma_i \) to denote the \( i \)-th largest singular value of \( W^* \). Then, \( \kappa \) is defined as \( \sigma_1 / \sigma_K \), and \( \gamma = \prod_{i=1}^K \sigma_i / \sigma_K \). Additionally, to avoid high dimensional tensors, the first order derivative of the empirical risk function is defined in the form of vectorized \( W \) as

\[
\nabla \hat{f}(W) = \begin{bmatrix} \partial f_1^T / \partial w_1 & \partial f_1^T / \partial w_2 & \cdots & \partial f_1^T / \partial w_K \end{bmatrix}^T \in \mathbb{R}^{dK}
\]

with \( W = [w_1, w_2, \cdots, w_K] \in \mathbb{R}^{d \times dK} \). Therefore, the second order derivative of the empirical risk function is in \( \mathbb{R}^{dK \times dK} \). Similar to \( \| \nabla^2 \hat{f} \|_2 \), the high order derivatives of the population risk functions are defined based on vectorized \( \hat{W} \) as well. In addition, without special descriptions, \( \alpha = [\alpha_1, \alpha_2, \cdots, \alpha_K]^T \) stands for any unit vector that in \( \mathbb{R}^{dK} \) with \( \alpha_j \in \mathbb{R}^d \). Therefore, we have

\[
\| \nabla^2 \hat{f} \|_2 = \max_{\alpha} \| \alpha^T \nabla^2 \hat{f} \alpha \|_2 = \max_{\alpha} \left( \sum_{j=1}^K \alpha_j^2 \frac{\partial \hat{f}}{\partial w_j} \right)^2.
\]
Finally, since we focus on order-wise analysis, some constant numbers will be ignored in the majority of the steps. In particular, we use \( h_1(z) \geq (\text{or } \leq) h_2(z) \) to denote there exists some positive constant \( C \) such that \( h_1(z) \geq (\text{or } \leq) C \cdot h_2(z) \) when \( z \in \mathbb{R} \) is sufficiently large.

### Table 3: Some Important Notations

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathcal{D} )</td>
<td>Labeled dataset with ( N ) number of samples;</td>
</tr>
<tr>
<td>( \tilde{\mathcal{D}} )</td>
<td>Unlabeled dataset with ( M ) number of samples;</td>
</tr>
<tr>
<td>( \mathcal{D}_t )</td>
<td>A subset of ( \mathcal{D} ) with ( N_t ) number of labeled data;</td>
</tr>
<tr>
<td>( \tilde{\mathcal{D}}_t )</td>
<td>A subset of ( \tilde{\mathcal{D}} ) with ( M_t ) number of unlabeled data;</td>
</tr>
<tr>
<td>( d )</td>
<td>Dimension of the input ( x ) or ( \tilde{x} );</td>
</tr>
<tr>
<td>( K )</td>
<td>Number of neurons in the hidden layer;</td>
</tr>
<tr>
<td>( W^* )</td>
<td>Weights of the ground truth model;</td>
</tr>
<tr>
<td>( W^{[p]} )</td>
<td>( W^{[p]} = pW^* + (1 - p)W^{(0,0)} );</td>
</tr>
<tr>
<td>( W^{(\ell,t)} )</td>
<td>Model returned by iterative self-training after ( t ) step mini-batch stochastic gradient descent at stage ( \ell ); ( W^{(0,0)} ) is the initial model;</td>
</tr>
<tr>
<td>( \hat{f}_{\mathcal{D}, \tilde{\mathcal{D}}}(\text{or } \hat{f}) )</td>
<td>The empirical risk function defined in [1];</td>
</tr>
<tr>
<td>( f(W; p) )</td>
<td>The population risk function defined in [17];</td>
</tr>
<tr>
<td>( \lambda )</td>
<td>The value of ( \lambda^2 / (\text{or } \lambda_{\tilde{\mathcal{D}}}) );</td>
</tr>
<tr>
<td>( \mu )</td>
<td>The value of ( \text{or } \mu_{\tilde{\mathcal{D}}} );</td>
</tr>
<tr>
<td>( \sigma_i )</td>
<td>The ( i )-th largest singular value of ( W^* );</td>
</tr>
<tr>
<td>( \kappa )</td>
<td>The value of ( \text{or } \sigma_1 / \sigma_K );</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>The value of ( \text{or } \prod_{i=1}^{K} \sigma_i / \sigma_K );</td>
</tr>
<tr>
<td>( q )</td>
<td>Some large constant in ( \mathbb{R}^+ );</td>
</tr>
</tbody>
</table>

## D Preliminary Lemmas

We will first start with some preliminary lemmas. As outlined at the beginning of the supplementary material, Lemma 1 illustrates the local convex region of the population risk function, and Lemma 2 explains the error bound between the population risk and empirical risk functions. Then, Lemma 3 describes the returned initial model \( W^{(0,0)} \) via tensor initialization method (Zhong et al., 2017) purely using labeled data. Next, Lemma 4 is the well known Weyl’s inequality in the matrix setting. Moreover, Lemma 5 is the concentration theorem for independent random matrices. The definitions of the sub-Gaussian and sub-exponential variables are summarized in Definitions 1 and 2. Lemmas 6 and 7 serve as the technical tools in bounding matrix norms under the framework of the confidence interval.

**Lemma 1.** Given any \( W \in \mathbb{R}^{d \times K} \), let \( p \) satisfy

\[
 p \lesssim \frac{\sigma_K}{\mu^2 K \cdot \| W - W^* \|_F}.
\]  

Then, we have

\[
 \frac{\lambda \rho(\delta) + \tilde{\lambda} \rho(\tilde{\delta})}{12K^2 \gamma K^2} \leq \nabla^2 f(W; p) \leq \frac{7(\lambda \delta^2 + \tilde{\lambda} \tilde{\delta}^2)}{K}.
\]
Lemma 2. Let $f$ and $\hat{f}$ be the functions defined in (17) and (1), respectively. Suppose the pseudo label is generated through (16) with weights $\hat{W}$. Then, we have

$$
\|\nabla f(W) - \nabla \hat{f}(W)\|_2 \leq \frac{\lambda^2}{K} \sqrt{\frac{d \log q}{N}} \cdot \|W - W^*\| + \frac{\tilde{\lambda} \hat{\delta}^2}{K} \sqrt{\frac{d \log q}{M}} \cdot \|W - \hat{W}\|_2
$$

with probability at least $1 - q^{-d}$.

Lemma 3 (Initialization, [Zhong et al. 2017]). Assuming the number of labeled data satisfies

$$
N \geq p^2 N^*
$$

for some large constant $q$ and $p \in \left[\frac{1}{K}, 1\right]$, the tensor initialization method, which is summarized in Appendix B, outputs $W^{(0,0)}$ such that

$$
\|W^{(0,0)} - W^*\| \leq \frac{\sigma_K}{\mu K}
$$

with probability at least $1 - q^{-d}$.

Lemma 4 (Weyl’s inequality, [Bhatia 2013]). Let $B = A + E$ be a matrix with dimension $m \times m$. Let $\lambda_i(B)$ and $\lambda_i(A)$ be the $i$-th largest eigenvalues of $B$ and $A$, respectively. Then, we have

$$
|\lambda_i(B) - \lambda_i(A)| \leq \|E\|_2, \quad \forall \quad i \in [m].
$$

Lemma 5 ([Tropp 2012], Theorem 1.6). Consider a finite sequence $\{Z_k\}$ of independent, random matrices with dimensions $d_1 \times d_2$. Assume that such random matrix satisfies

$$
\mathbb{E}(Z_k) = 0 \quad \text{and} \quad \|Z_k\| \leq R \quad \text{almost surely}.
$$

Define

$$
\delta^2 := \max \left\{ \| \sum_k \mathbb{E}(Z_k Z_k^\top) \|, \| \sum_k \mathbb{E}(Z_k^\top Z_k) \| \right\}.
$$

Then for all $t \geq 0$, we have

$$
\text{Prob}\left\{ \left\| \sum_k Z_k \right\| \geq t \right\} \leq (d_1 + d_2) \exp \left( \frac{-t^2/2}{\delta^2 + R t/3} \right).
$$

Definition 1 (Definition 5.7, [Vershynin 2010]). A random variable $X$ is called a sub-Gaussian random variable if it satisfies

$$
(\mathbb{E}|X|^p)^{1/p} \leq c_1 \sqrt{p}
$$

for all $p \geq 1$ and some constant $c_1 > 0$. In addition, we have

$$
\mathbb{E}e^{s(X - \mathbb{E}X)} \leq e^{c_2 \|X\|^2_{\psi_2} s^2}
$$

for all $s \in \mathbb{R}$ and some constant $c_2 > 0$, where $\|X\|_{\psi_2}$ is the sub-Gaussian norm of $X$ defined as

$$
\|X\|_{\psi_2} = \sup_{p \geq 1} p^{-1/2} (\mathbb{E}|X|^p)^{1/p}.
$$

Moreover, a random vector $X \in \mathbb{R}^d$ belongs to the sub-Gaussian distribution if one-dimensional marginal $\alpha^T X$ is sub-Gaussian for any $\alpha \in \mathbb{R}^d$, and the sub-Gaussian norm of $X$ is defined as

$$
\|X\|_{\psi_2} = \sup_{\|\alpha\|_2 = 1} \|\alpha^T X\|_{\psi_2}.
$$

Definition 2 (Definition 5.13, [Vershynin 2010]). A random variable $X$ is called a sub-exponential random variable if it satisfies

$$
(\mathbb{E}|X|^p)^{1/p} \leq c_3 p
$$

for all $p \geq 1$ and some constant $c_3 > 0$. In addition, we have

$$
\mathbb{E}e^{s(X - \mathbb{E}X)} \leq e^{c_4 \|X\|^2_{\psi_1} s^2}
$$

for $s \leq 1/\|X\|_{\psi_1}$ and some constant $c_4 > 0$, where $\|X\|_{\psi_1}$ is the sub-exponential norm of $X$ defined as

$$
\|X\|_{\psi_1} = \sup_{p \geq 1} p^{-1} (\mathbb{E}|X|^p)^{1/p}.
$$
Lemma 6 (Lemma 5.2, [Vershynin 2010]). Let \( B(0, 1) \subset \{ \alpha ||\alpha||_2 = 1, \alpha \in \mathbb{R}^d \} \) denote a unit ball in \( \mathbb{R}^d \). Then, a subset \( S_\xi \) is called a \( \xi \)-net of \( B(0, 1) \) if every point \( z \in B(0, 1) \) can be approximated to within \( \xi \) by some point \( \alpha \in B(0, 1) \), i.e., \( \|z - \alpha\|_2 \leq \xi \). Then the minimal cardinality of a \( \xi \)-net \( S_\xi \) satisfies
\[
|S_\xi| \leq (1 + 2/\xi)^d. 
\]  

Lemma 7 (Lemma 5.3, [Vershynin 2010]). Let \( A \) be an \( d_1 \times d_2 \) matrix, and let \( S_\xi(d) \) be a \( \xi \)-net of \( B(0, 1) \) in \( \mathbb{R}^d \) for some \( \xi \in (0, 1) \). Then
\[
\|A\|_2 \leq (1 - \xi)^{-1} \max_{\alpha_1, \alpha_2 \in S_\xi(d_1), \alpha_2 \in S_\xi(d_2)} |\alpha_1^T A \alpha_2|. 
\]

Lemma 8 (Mean Value Theorem). Let \( U \subset \mathbb{R}^{n_1} \) be open and \( f : U \rightarrow \mathbb{R}^{n_2} \) be continuously differentiable, and \( x \in U \), \( h \in \mathbb{R}^{n_1} \) vectors such that the line segment \( x + t h, 0 \leq t \leq 1 \) remains in \( U \). Then we have:
\[
f(x + h) - f(x) = \left( \int_0^1 \nabla f(x + th)dt \right) \cdot h,
\]
where \( \nabla f \) denotes the Jacobian matrix of \( f \).

E PROOF OF THEOREM 2

With \( p = 1 \) in (17), the population risk function is reduced as
\[
f(W) = \frac{\lambda}{2} \mathbb{E}_x (y - g(W; x)) + \frac{\lambda}{2} \mathbb{E}_{\tilde{x}} (\tilde{y} - g(W; \tilde{x})),
\]
where \( y = g(W^*; x) \) with \( x \sim \mathcal{N}(0, \delta^2 I) \) and \( \tilde{y} = g(W^*; \tilde{x}) \) with \( \tilde{x} \sim \mathcal{N}(0, \tilde{\delta}^2 I) \). In fact, (32) can be viewed as the expectation of the empirical risk function in (1) given \( \tilde{y}_m = g(W^*; \tilde{x}_m) \). Moreover, the ground-truth model \( W^* \) is the global optimal to (32), as well. Lemmas 8 and 10 are the special case of Lemmas 1 and 2 with \( p = 1 \). The proof of Theorem 2 is followed by the presentation of the two lemmas.

The main idea in proving Theorem 2 is to characterize the gradient descent term by the MVT in Lemma 8 as shown in (36) and (37). The IVT is not directly applied in the empirical risk function because of its non-smoothness. However, the population risk functions defined in (17) and (32), which are the expectations over the Gaussian variables, are smooth. Then, as the distance \( \|\nabla f(W) - \nabla f(W^*)\|_F \) is upper bounded by a linear function of \( \|W - W^*\|_F \) as shown in (47), we can establish the connection between \( \|W^{(\ell, t+1)} - W^*\|_F \) and \( \|W^{(\ell, t)} - W^*\|_F \) as shown in (50). Finally, by mathematical induction over \( \ell \) and \( t \), one can characterize \( \|W^{(L, 0)} - W^*\|_F \) by \( \|W^{(0, 0)} - W^*\|_F \) as shown in (52), which completes the whole proof.

Lemma 9 (Lemma 1 with \( p = 1 \)). Let \( f \) and \( \hat{f} \) be the functions defined in (32) and (1), respectively. Then, for any \( W \) that satisfies,
\[
\|W - W^*\|_F \leq \frac{\sigma_K}{\mu^2 K}, 
\]
we have
\[
\frac{\lambda \rho(\delta) + \tilde{\lambda} \rho(\tilde{\delta})}{12n^2 \gamma K^2} \leq \nabla^2 f(W) \leq \frac{7(\lambda \delta^2 + \tilde{\lambda} \tilde{\delta}^2)}{K}. 
\]

Lemma 10 (Lemma 2 with \( p = 1 \)). Let \( f \) and \( \hat{f} \) be the functions defined in (32) and (1), respectively. Suppose the pseudo label is generated through (16) with weights \( \hat{W} \). Then, we have
\[
\|\nabla f(W) - \nabla \hat{f}(W)\|_2 \leq \left( \frac{\lambda \delta^2}{K} \sqrt{\frac{d \log q}{N}} + \frac{1 - \lambda}{K} \sqrt{\frac{d \log q}{M}} \right) \cdot \|W - W^*\|_2
\]
\[
+ \frac{(1 - \lambda) \delta^2}{K} \left( \sqrt{\frac{d \log q}{M}} + \frac{1}{2} \right) \cdot \|\hat{W} - W^*\|_2
\]
with probability at least \( 1 - q^{-d} \).
Proof of Theorem[2] From Algorithm[1] in the $\ell$-th outer loop, we have
\[
W^{(\ell,t+1)} &= W^{(\ell,t)} - \eta \nabla \tilde{f}_{D_t, \beta_t}(W^{(\ell,t)}) + \beta (W^{(\ell,t)} - W^{(\ell,t-1)}) \\
&= W^{(\ell,t)} - \eta \nabla f(W^{(\ell,t)}) + \beta (W^{(\ell,t)} - W^{(\ell,t-1)}) \\
&\quad + \eta \cdot (\nabla f(W^{(\ell,t)}) - \nabla \tilde{f}_{D_t, \beta_t}(W^{(\ell,t)})).
\]
(36)

Since $\nabla f$ is a smooth function and $W^*$ is a local (global) optimal to $f$, then we have
\[
\nabla f(W^{(\ell,t)}) = \nabla f(W^{(\ell,t)}) - \nabla f(W^*)
\]
\[
= \int_0^1 \nabla^2 f\left(W^{(\ell,t)} + u \cdot (W^{(\ell,t)} - W^*)\right) du \cdot (W^{(\ell,t)} - W^*),
\]
(37)

where the last equality comes from MVT in Lemma[8] For notational convenience, we use $H^{(\ell,t)}$ to denote the integration as
\[
H^{(\ell,t)} := \int_0^1 \nabla^2 f\left(W^{(\ell,t)} + u \cdot (W^{(\ell,t)} - W^*)\right) du.
\]
(38)

Then, we have
\[
\begin{bmatrix}
W^{(\ell,t+1)} - W^* \\
W^{(\ell,t)} - W^*
\end{bmatrix} =
\begin{bmatrix}
I - \eta H^{(\ell,t)} & \beta I \\
I & 0
\end{bmatrix}
\begin{bmatrix}
W^{(\ell,t)} - W^* \\
W^{(\ell,t-1)} - W^*
\end{bmatrix}
\]
\[
+ \eta \begin{bmatrix}
\nabla f(W^{(\ell,t)}) - \nabla \tilde{f}_{D_t, \beta_t}(W^{(\ell,t)}) \\
0
\end{bmatrix}.
\]
(39)

Let $H^{(\ell,t)} = SAS^T$ be the eigen-decomposition of $H^{(\ell,t)}$. Then, we define
\[
A(\beta) :=
\begin{bmatrix}
S^T & 0 \\
0 & S^T
\end{bmatrix} A(\beta)
\begin{bmatrix}
S & 0 \\
0 & S
\end{bmatrix} =
\begin{bmatrix}
I - \eta A + \beta I & \beta I \\
I & 0
\end{bmatrix}.
\]
(40)

Since $\begin{bmatrix}
S & 0 \\
0 & S
\end{bmatrix} \begin{bmatrix}
S^T & 0 \\
0 & S^T
\end{bmatrix} = \begin{bmatrix}
I & 0 \\
0 & I
\end{bmatrix}$, we know $A(\beta)$ and $\begin{bmatrix}
I - \eta A + \beta I & \beta I \\
I & 0
\end{bmatrix}$ share the same eigenvalues. Let $\gamma_i(\beta)$ be the $i$-th eigenvalue of $\nabla^2 f(\tilde{w}^{(\ell)})$, then the corresponding $i$-th eigenvalue of (40), denoted by $\gamma_i^{(\beta)}$, satisfies
\[
(\gamma_i^{(\beta)})^2 - (1 - \eta \gamma_i^{(\beta)} + \beta) \gamma_i^{(\beta)} + \beta = 0.
\]
(41)

By simple calculation, we have
\[
|\gamma_i^{(\beta)}| = \begin{cases}
\sqrt{\beta}, & \beta \geq (1 - \sqrt{\gamma_i^{(\beta)}})^2, \\
\frac{1}{2} \left(1 - \eta \gamma_i^{(\beta)} + \beta + \sqrt{(1 - \eta \gamma_i^{(\beta)} + \beta)^2 - 4\beta}\right), & \text{otherwise.}
\end{cases}
\]
(42)

Specifically, we have
\[
\gamma_i^{(A)}(0) > \gamma_i^{(A)}(\beta), \quad \text{for} \quad \forall \beta \in (0, (1 - \eta \gamma_i^{(A)})^2),
\]
(43)

and $\gamma_i^{(A)}$ achieves the minimum $\gamma_i^{(A)*} = 1 - \sqrt{\eta \gamma_i^{(A)}}$ when $\beta = \left(1 - \sqrt{\eta \gamma_i^{(A)}}\right)^2$. From Lemma[2] for any $a \in \mathbb{R}^d$ with $||a||_2 = 1$, we have
\[
a^T \nabla f(W^{(\ell,t)}) a = \int_0^1 a^T \nabla^2 f\left(W^{(\ell,t)} + u \cdot (W^{(\ell,t)} - W^*)\right) a du \leq \int_0^1 \gamma_{\text{max}} ||a||_2^2 du = \gamma_{\text{max}},
\]
\[
a^T \nabla f(W^{(\ell,t)}) a = \int_0^1 a^T \nabla^2 f\left(W^{(\ell,t)} + u \cdot (W^{(\ell,t)} - W^*)\right) a du \geq \int_0^1 \gamma_{\text{min}} ||a||_2^2 du = \gamma_{\text{min}},
\]
(44)
where $\gamma_{\text{max}} = \frac{7(\lambda \delta^2 + \tilde{\lambda} \delta^2)}{K}$, and $\gamma_{\text{min}} = \frac{\lambda \rho(\delta) + \tilde{\lambda} \rho(\delta)}{12 \kappa^2 \gamma K^2}$. Therefore, we have

$$\gamma_{\text{min}} = \frac{\lambda \rho(\delta) + \tilde{\lambda} \rho(\delta)}{12 \kappa^2 \gamma K^2}, \quad \text{and} \quad \gamma_{\text{max}} = \frac{7(\lambda \delta^2 + \tilde{\lambda} \delta^2)}{K}. \quad (45)$$

Thus, we can select $\eta = \left(\frac{1}{\sqrt{\gamma_{\text{max}}^2 + \gamma_{\text{min}}^2}}\right)^2$, and $\|A(\beta)\|_2$ can be bounded by

$$\min_{\beta} \|A(\beta)\|_2 \leq 1 - \sqrt{\frac{(\lambda \rho(\delta) + \tilde{\lambda} \rho(\delta))}{12 \kappa^2 \gamma K^2}} / (2 \cdot \frac{7(\lambda \delta^2 + \tilde{\lambda} \delta^2)}{K})$$

$$= 1 - \frac{\mu(\delta, \tilde{\delta})}{\sqrt{168 \kappa^2 \gamma K}}, \quad (46)$$

where $\mu(\delta, \tilde{\delta}) = \left(\frac{\lambda \rho(\delta) + \tilde{\lambda} \rho(\delta)}{\lambda \delta^2 + \tilde{\lambda} \delta^2}\right)^{1/2}$.

From Lemma [16], we have

$$\|\nabla f(W^{(\ell, t)}) - \nabla \hat{f}(W^{(\ell, t)})\|_2 = \left(\frac{\lambda \delta^2}{K} \sqrt{\frac{d \log q}{N_t}} + \tilde{\lambda} \delta^2 \sqrt{\frac{d \log q}{M_t}}\right) \cdot \|W^{(\ell, t)} - W^*\|_2$$

$$+ \frac{\tilde{\lambda} \delta^2}{K} \left(\sqrt{\frac{d \log q}{M_t}} + \frac{1}{2}\right) \cdot \|W^{(\ell, 0)} - W^*\|_2. \quad (47)$$

Given $\varepsilon > 0$ and $\tilde{\varepsilon} > 0$ with $\varepsilon + \tilde{\varepsilon} < 1$, let

$$\eta \cdot \frac{\lambda \delta^2}{K} \sqrt{\frac{d \log q}{N_t}} \leq \frac{\varepsilon \mu(\delta, \tilde{\delta})}{\sqrt{168 \kappa^2 \gamma K}}.$$  

and

$$\eta \cdot \frac{\tilde{\lambda} \delta^2}{K} \sqrt{\frac{d \log q}{M_t}} \leq \frac{\tilde{\varepsilon} \mu(\delta, \tilde{\delta})}{\sqrt{168 \kappa^2 \gamma K}}, \quad (48)$$

where we need

$$N_t \geq \varepsilon^{-2} \mu^{-2} \left(\frac{\lambda \delta^2}{\lambda \delta^2 + \tilde{\lambda} \delta^2}\right)^2 \kappa^2 \gamma K^3 d \log q,$$

and

$$M_t \geq \tilde{\varepsilon}^{-2} \mu^{-2} \left(\frac{\tilde{\lambda} \delta^2}{\lambda \delta^2 + \tilde{\lambda} \delta^2}\right)^2 \kappa^2 \gamma K^3 d \log q. \quad (49)$$

Therefore, from [46], [47], and [48], we have

$$\|W^{(\ell, t+1)} - W^*\|_2 \leq \left(1 - \frac{(1 - \varepsilon - \tilde{\varepsilon}) \mu(\delta, \tilde{\delta})}{\sqrt{168 \kappa^2 \gamma K}}\right) \|W^{(\ell, t)} - W^*\|_2 + \eta \cdot \frac{\tilde{\lambda} \delta^2}{K} \left(\sqrt{\frac{d \log q}{M_t}} + \frac{1}{2}\right) \cdot \|W^{(\ell, 0)} - W^*\|_2$$

$$\leq \left(1 - \frac{(1 - \varepsilon - \tilde{\varepsilon}) \mu(\delta, \tilde{\delta})}{\sqrt{168 \kappa^2 \gamma K}}\right) \|W^{(\ell, t)} - W^*\|_2 + \eta \cdot \frac{\tilde{\lambda} \delta^2}{K} \|W^{(\ell, 0)} - W^*\|_2$$

$$\leq \left(1 - \frac{(1 - \varepsilon - \tilde{\varepsilon}) \mu(\delta, \tilde{\delta})}{\sqrt{168 \kappa^2 \gamma K}}\right) \|W^{(\ell, t)} - W^*\|_2 + \frac{\sqrt{K}}{14(\lambda \delta^2 + \tilde{\lambda} \delta^2)} \cdot \frac{\tilde{\lambda} \delta^2}{K} \|W^{(\ell, 0)} - W^*\|_2 \quad (50)$$

when $M \geq 4d \log q$. By mathematical induction on [50] over $t$, we have

$$\|W^{(\ell, t)} - W^*\|_2 \leq \left(1 - \frac{(1 - \varepsilon - \tilde{\varepsilon}) \mu(\delta, \tilde{\delta})}{\sqrt{168 \kappa^2 \gamma K}}\right)^t \cdot \|W^{(\ell, 0)} - W^*\|_2$$

$$+ \frac{\sqrt{168 \kappa^2 \gamma K}}{(1 - \varepsilon - \tilde{\varepsilon}) \mu(\lambda \delta^2 + \tilde{\lambda} \delta^2)} \cdot \frac{\tilde{\lambda} \delta^2}{K} \|W^{(\ell, 0)} - W^*\|_2 \quad (51)$$

$$\leq \left(1 - \frac{(1 - \varepsilon - \tilde{\varepsilon}) \mu(\delta, \tilde{\delta})}{\sqrt{168 \kappa^2 \gamma K}}\right)^t + \frac{\sqrt{K \lambda \delta^2}}{(1 - \varepsilon - \tilde{\varepsilon}) \mu(\lambda \delta^2 + \tilde{\lambda} \delta^2)} \cdot \|W^{(\ell, 0)} - W^*\|_2$$
By mathematical induction on (51) over \( \ell \), we have
\[
\|W^{(\ell,T)} - W^*\|_2 \leq \left[ \left( 1 - \frac{(1 - \varepsilon - \tilde{\varepsilon})\mu}{\sqrt{168\kappa^2\gamma K}} \right) T + \frac{\sqrt{\kappa^2\gamma}\lambda\hat{\delta}^2}{(1 - \varepsilon - \tilde{\varepsilon})\mu(\lambda\delta^2 + \lambda\tilde{\delta}^2)} \right] \cdot \|W^{(0,0)} - W^*\|_2
\]
(52)

\[\Box\]

\[\text{Proof of Theorem 3}\]

Instead of proving Theorem 1, we turn to prove a stronger version, as shown in Theorem 3. One can verify that Theorem 1 is a special case of Theorem 3 by selecting \( \lambda \) in the order of \( p \) and \( \varepsilon \) in the order of \( (2p - 1) \).

The major idea in proving Theorem 3 is similar to that of Theorem 2. The first step is to characterize the gradient descent term on the population risk function by the MVT in Lemma 8 as shown in (64). Compared with proving Theorem 2 where the induction over (58) and (59). Then, the connection between \( \|W^{(\ell+1,0)} - W^{[p]}\|_F \) and \( \|W^{(\ell,0)} - W^{[p]}\|_F \) are characterized in (64). Compared with proving Theorem 2 where the induction over \( \ell \) holds naturally with large size of labeled data, the induction over \( \ell \) requires a proper value of \( p \) as shown in (69). By induction over \( \ell \) on (64), the relative error \( \|W^{(L,0)} - W^{[p]}\|_F \) can be characterized by \( \|W^{(\ell,0)} - W^{[p]}\|_F \) as shown in (71).

**Theorem 3.** Suppose the initialization \( W^{(0,0)} \) satisfies with
\[
|p - \hat{\lambda}| \leq \frac{2(1 - \tilde{\varepsilon})p - 1}{\mu\sqrt{K}}
\]
(53)
for some constant \( \tilde{\varepsilon} \in (0, 1/2) \), where
\[
\hat{\lambda} := \frac{\lambda\delta^2}{\lambda\delta^2 + \lambda\tilde{\delta}^2} = \left( \frac{N}{\kappa^2\gamma K^3\mu^2 d \log q} \right) ^{1/2}
\]
(54)
and
\[
\mu = \mu(\delta, \tilde{\delta}) = \frac{\lambda\delta^2 + \lambda\tilde{\delta}^2}{\lambda\rho(\delta) + \lambda\rho(\tilde{\delta})}.
\]
(55)

Then, if the number of samples in \( \tilde{D} \) further satisfies
\[
M \geq \varepsilon^{-2}\kappa^2\gamma^2\mu^2(1 - \hat{\lambda})^2K^3d\log q,
\]
(56)
the iterates \( \{W^{(\ell,t)}\}_{\ell,t=0}^{L,T} \) converge to \( W^{[p]} \) with \( p \) satisfies (53) as
\[
\lim_{T \to \infty} \|W^{(\ell,T)} - W^{[p]}\|_2 \leq \frac{1}{1 - \tilde{\varepsilon}} \cdot \left( 1 - p^* + \mu\sqrt{K}|(\hat{\lambda} - p^*)| \right) \cdot \|W^{(0,0)} - W^*\|_2 + \frac{\varepsilon}{1 - \tilde{\varepsilon}} \cdot \|W^{(\ell,0)} - W^{[p]}\|_2,
\]
(57)
with probability at least 1 - \( q^{-d} \).

**Proof of Theorem 3.** From Algorithm 1 in the \( \ell \)-th outer loop, we have
\[
W^{(\ell, t+1)} = W^{(\ell, t)} - \eta\nabla f_{D_t, \hat{D}_t}(W^{(\ell, t)}) + \beta(W^{(\ell, t)} - W^{(\ell, t-1)})
\]
\[
= W^{(\ell, t)} - \eta\nabla f(W^{(\ell, t)}) + \beta(W^{(\ell, t)} - W^{(\ell, t-1)}) + \eta \cdot \left( \nabla f(W^{(\ell, t)}) - \nabla f_{D_t, \hat{D}_t}(W^{(\ell, t)}) \right)
\]
(58)

Since \( \nabla f \) is a smooth function and \( W^{[p]} \) is a local (global) optimal to \( f \), then we have
\[
\nabla f(W^{(\ell, t)}) = \nabla f(W^{(\ell, t)}) - \nabla f(W^{[p]})
\]
\[
= \int_0^1 \nabla^2 f(W^{(\ell, t)} + u \cdot (W^{(\ell, t)} - W^{[p]})) du \cdot (W^{(\ell, t)} - W^{[p]}),
\]
(59)
where the last equality comes from Lemma 8.

Similar to the proof of Theorem 2 we have
\[
\|W^{(\ell+1)} - W^p\|_2 \leq \|A(\beta)\|_2 \cdot \|W^{(\ell)} - W^p\|_2 + \eta \cdot \|\nabla f(W^{(\ell)}) - \nabla \hat{f}_{D_1, D_2}(W^{(\ell)})\|_2. \tag{60}
\]

From Lemma 2 we have
\[
\|\nabla f(W^{(\ell)}) - \nabla \hat{f}(W^{(\ell)})\|_2 \leq \frac{\lambda^2}{K} \sqrt{\frac{d \log q}{N_t}} \cdot \|W^{(\ell)} - W^p\| + \frac{\bar{\lambda} \delta^2}{K} \sqrt{\frac{d \log q}{M_t}} \cdot \|W^p - W\|_2 + \frac{\lambda \delta^2 \cdot (W^{(0,0)} - W^p)}{K} \cdot \|W^{(0,0)} - W^p\|_2,
\]
\[
\leq \frac{\lambda^2}{K} \sqrt{\frac{d \log q}{N_t}} \cdot \|W^{(\ell)} - W^p\| + \frac{\bar{\lambda} \delta^2}{K} \sqrt{\frac{d \log q}{M_t}} \cdot \|W^{(\ell)} - W^p\|_2 + \frac{\lambda \delta^2 \cdot (1 - p) - \bar{\lambda} \delta^2 \cdot p}{K} \cdot \|W^{(0,0)} - W^p\|_2.
\]

When \(\ell = 0\), following the similar steps from (41) to (46), we have
\[
\|\nabla f(W^{(\ell)}) - \nabla \hat{f}(W^{(\ell)})\|_2 \leq \frac{\lambda^2}{K} \sqrt{\frac{d \log q}{N_t}} \cdot \|W^{(\ell)} - W^p\| + \frac{\bar{\lambda} \delta^2}{K} \sqrt{\frac{d \log q}{M_t}} \cdot \|W^{(\ell)} - W^p\|_2 + \frac{\lambda \delta^2 \cdot (1 - p) - \bar{\lambda} \delta^2 \cdot p}{K} \cdot \|W^{(0,0)} - W^p\|_2.
\]
\[
\leq \left(1 - \frac{1 - \tilde{\varepsilon}}{\mu(\delta, \nu) \sqrt{154 K^2 \gamma K}}\right) \cdot \|W^{(\ell)} - W^p\|_2 + \eta \cdot \left(\frac{\lambda \delta^2 (1 - p)}{K} \sqrt{\frac{d \log q}{N_t}} + \frac{\lambda \delta^2 \cdot (1 - p) - \bar{\lambda} \delta^2 \cdot p}{K}\right) \cdot \|W^{(0,0)} - W^p\|_2 + \eta \cdot \frac{\varepsilon \bar{\lambda} \delta^2 \cdot p}{K} \cdot \sqrt{\frac{d \log q}{M_t}} \cdot \|W^{(0,0)} - W^p\|_2.
\]
\[
\leq \frac{\mu \sqrt{154 K^2 \gamma K}}{1 - \tilde{\varepsilon}} \cdot \eta \cdot \left(\frac{\lambda \delta^2 (1 - p)}{K} \sqrt{\frac{d \log q}{N_t}} + \frac{\lambda \delta^2 \cdot (1 - p) - \bar{\lambda} \delta^2 \cdot p}{K}\right) \cdot \|W^{(0,0)} - W^p\|_2 + \frac{\varepsilon \bar{\lambda} \delta^2 \cdot p}{K} \cdot \sqrt{\frac{d \log q}{M_t}} \cdot \|W^{(0,0)} - W^p\|_2
\]
\[
\leq \frac{1}{1 - \tilde{\varepsilon}} \cdot \left(1 - p + \sqrt{K} \cdot (1 - p) \mu \tilde{\lambda} - p \mu (1 - \tilde{\lambda})\right) \cdot \|W^{(0,0)} - W^p\|_2 + \frac{\varepsilon \bar{p}}{1 - \tilde{\varepsilon}} \cdot \|W^{(0,0)} - W^p\|_2,
\]
\[
= \frac{1}{1 - \tilde{\varepsilon}} \cdot \left(1 - p + \mu \sqrt{K} \tilde{\lambda} - p \right) \cdot \|W^{(0,0)} - W^p\|_2 + \frac{\varepsilon \bar{p}}{1 - \tilde{\varepsilon}} \cdot \|W^{(0,0)} - W^p\|_2,
\]
where \(\tilde{\lambda} = \frac{\lambda \delta^2}{\lambda \delta^2 + \bar{\lambda} \delta^2}\).
To guarantee the convergence in the outer loop, we require
\[ \lim_{T \to \infty} \| W^{(\ell,T)} - W^p \|_2 \leq \| W^{(0,0)} - W^p \|_2 = p \| W^{(0,0)} - W^* \|_2, \]
and
\[ \lim_{T \to \infty} \| W^{(\ell,T)} - W^* \|_2 \leq \| W^{(0,0)} - W^* \|_2. \] (65)
Since we have
\[ \| W^{(\ell,T)} - W^p \|_2 \leq \| W^{(\ell,T)} - W^* \|_2 + \| W^* - W^p \|_2 \]
\[ = \| W^{(\ell,T)} - W^* \|_2 + (1 - p) \cdot \| W^* - W^{(0,0)} \|_2, \] (66)
it is clear that (65) holds if and only if
\[ \frac{1}{1 - \bar{\varepsilon}} \cdot \left( 1 - p + \bar{\varepsilon} p + \mu \sqrt{K} | \hat{\lambda} - p | \right) + 1 - p \leq 1. \] (67)
To guarantee the iterates strictly converges to the desired point, we let
\[ \frac{1}{1 - \bar{\varepsilon}} \cdot \left( 1 - p + \bar{\varepsilon} p + \mu \sqrt{K} | \hat{\lambda} - p | \right) + 1 - p \leq 1 - \frac{1}{C} \] (68)
for some larger constant $C$, which is equivalent to
\[ | p - \hat{\lambda} | \leq \frac{2(1 - \bar{\varepsilon})p - 1}{\mu \sqrt{K}}. \] (69)
To make the bound in (69) meaningful, we need
\[ p \geq \frac{1}{2(1 - \bar{\varepsilon})}. \] (70)
When $\ell > 1$, following similar steps in (64), we have
\[ \lim_{T \to \infty} \| W^{(\ell,T)} - W^p \|_2 \leq \frac{1}{1 - \bar{\varepsilon}} \cdot \left( 1 - p + \mu \sqrt{K} | \hat{\lambda} - p | \right) \cdot \| W^{(0,0)} - W^* \|_2 + \frac{\bar{\varepsilon} p}{1 - \bar{\varepsilon}} \cdot \| W^{(\ell,0)} - W^p \|_2, \] (71)
Given (69) holds, from (71), we have
\[ \lim_{L \to \infty, T \to \infty} \| W^{(L,T)} - W^p \|_2 \leq \frac{1}{1 - \bar{\varepsilon}} \cdot \left( 1 - p + \mu \sqrt{K} | \hat{\lambda} - p | \right) \cdot \| W^{(0,0)} - W^* \|_2 \] (72)
\[ \leq \frac{1}{1 - \bar{\varepsilon}} \cdot \left( 1 - p + \mu \sqrt{K} | \hat{\lambda} - p | \right) \cdot \| W^{(0,0)} - W^* \|_2. \]

\[ \square \]

### G Definition and relative proofs of $\rho$

In this section, the formal definition of $\rho$ is included in Definition 3 and a corresponding claim about $\rho$ is summarized in Lemma 11. One can quickly check that the ReLU activation function satisfies the conditions in Lemma 11.

The major idea in proving Lemma 11 is to show $H_\ell(\delta)$ and $J_\ell(\delta)$ in Definition 3 are in the order of $\delta^r$ when $\delta$ is small.

**Definition 3.** Let $H_\ell(\delta) = E_{Z \sim N(0, \delta^2)}(\phi'(\sigma_K z) z^r)$ and $J_\ell(\delta) = E_{Z \sim N(0, \delta^2)}(\phi'^2(\sigma_K z) z^r)$. Then, $\rho = \rho(\delta)$ is defined as
\[ \rho(\delta) = \min \left\{ J_0(\delta) - H_0^2(\delta) - H_1^2(\delta), J_2(\delta) - H_1^2(\delta) - H_2^2(\delta), H_0(\delta) \cdot H_2(\delta) - H_1^2(\delta) \right\}, \] (73)
where $\sigma_K$ is the minimal singular value of $W^*$. 

24
Lemma 11 (Order analysis of ρ). If ρ(δ) > 0 for δ ∈ (0, ξ) for some positive constant ξ and the sub-gradient of ρ(δ) at 0 can be non-zero, then ρ(δ) = Θ(δ^2) when δ → 0+. Typically, for ReLU activation function, μ in [5] is a fixed constant for all δ, δ ≤ 1.

Proof of Lemma 11. From Definition 3, we know that H_0(δ) = E_{z \sim N(0, δ^2)} φ'(σ_K z) z^r. Suppose we have H_r(δ) = Θ(δ^r) and J_r(δ) = Θ(δ^r), then from (73) we have
\[
\begin{align*}
J_0(δ) - H_0^2(δ) - H_1^2(δ) & \in Θ(1) - Θ(δ^2), \\
J_2(δ) - H_2^2(δ) - H_1^2(δ) & \in Θ(δ^2), \\
H_0(δ) \cdot H_2(δ) - H_1^2(δ) & \in Θ(δ^2) - Θ(δ^4).
\end{align*}
\]
Because ρ is a continuous function with ρ(z) > 0 for some z > 0. Therefore, ρ ≠ J_0(δ) - H_0^2(δ) - H_1^2(δ) when δ → 0+, otherwise ρ(z) < 0 for any z > 0. When δ → 0+, both J_2(δ) - H_2^2(δ) and H_0(δ) \cdot H_2(δ) - H_1^2(δ) are in the order of δ^2, which indicates that μ is a fixed constant when both δ and δ̂ are close to 0. In addition, J_2(δ) - H_2^2(δ) - H_1^2(δ) goes to +∞ while both J_0(δ) - H_0^2(δ) - H_1^2(δ) and H_0(δ) \cdot H_2(δ) - H_1^2(δ) go to −∞ when δ → +∞. Therefore, with a large enough δ, we have
\[
ρ(δ) \in Θ(δ^2) - Θ(δ^4) \text{ or } Θ(1) - Θ(δ^2),
\]
which indicates that μ is a strictly decreasing function when δ and δ̂ are large enough.

Next, we provide the conditions that guarantee H_r(δ) = Θ(δ^r) hold, and the relative proof for J_r(δ) can be derived accordingly following the similar steps as well. From Definition 3 we have
\[
\begin{align*}
\lim_{δ → 0^+} \frac{H_r(δ)}{δ^r} & = \lim_{δ → 0^+} \frac{1}{δ^r} \int_{-∞}^{+∞} φ'(σ_K z) (\frac{z}{δ})^r \frac{1}{\sqrt{2πδ}} e^{-\frac{z^2}{2δ^2}} dz \\
& \xrightarrow{(a)} \lim_{δ → 0^+} \int_{-∞}^{+∞} φ'(σ_K δt) \frac{t^r}{\sqrt{2π}} e^{-\frac{t^2}{2}} dt \\
& = \lim_{δ → 0^+} \int_{-∞}^{0^-} φ'(σ_K δt) \frac{t^r}{\sqrt{2π}} e^{-\frac{t^2}{2}} dt + \lim_{δ → 0^+} \int_{0^+}^{+∞} φ'(σ_K δt) \frac{t^r}{\sqrt{2π}} e^{-\frac{t^2}{2}} dt \\
& = φ'(0^-) \int_{-∞}^{0^-} \frac{t^r}{\sqrt{2π}} e^{-\frac{t^2}{2}} dt + φ'(0^+) \int_{0^+}^{+∞} \frac{t^r}{\sqrt{2π}} e^{-\frac{t^2}{2}} dt,
\end{align*}
\]
where equality (a) holds by letting t = \frac{z}{δ}. It is easy to verify that
\[
\int_{0^+}^{+∞} \frac{t^r}{\sqrt{2π}} e^{-\frac{t^2}{2}} dt = (-1)^r \int_{-∞}^{0^-} \frac{t^r}{\sqrt{2π}} e^{-\frac{t^2}{2}} dt,
\]
and both are bounded for a fixed r. Thus, as long as either φ'(0^-) or φ'(0^+) is non-zero, we have H_r(δ) = Θ(δ^r) when δ → 0^+.

If φ has bounded gradient as |φ'| ≤ C_φ for some positive constant C_φ. Then, we have
\[
\begin{align*}
\left| \frac{H_r(δ)}{δ^r} \right| & = \left| \int_{-∞}^{+∞} φ'(σ_K z) (\frac{z}{δ})^r \frac{1}{\sqrt{2πδ}} e^{-\frac{z^2}{2δ^2}} dz \right| \\
& = \left| \int_{-∞}^{+∞} φ'(σ_K δt) \frac{t^r}{\sqrt{2π}} e^{-\frac{t^2}{2}} dt \right| \\
& ≤ C_φ \cdot \left| \int_{-∞}^{+∞} \frac{t^r}{\sqrt{2π}} e^{-\frac{t^2}{2}} dt \right| \\
\end{align*}
\]
Therefore, we have H_r(δ) = O(δ^r) for all δ > 0 when φ has bounded gradient.

Typically, for ReLU function, one can directly calculate that H_r(δ) = δ^r for δ ∈ R, and ρ(δ) = C_δ^2 when δ ≤ 1 for some constant C = 0.091. Then, it is easy to check that μ is a constant when δ, δ̂ ≤ 1. □
H PROOF OF PRELIMINARY LEMMAS

H.1 PROOF OF LEMMA 1

The eigenvalues of $\nabla^2 f(\cdot; p)$ at any fixed point $W$ can be bounded in the form of (80) by Weyl’s inequality (Lemma 4). Therefore, the primary technical challenge lies in bounding $\|\nabla^2 f(W; p) - \nabla^2 f(W[p]; p)\|_2$, which is summarized in Lemma 12. Lemma 13 provides the exact calculation of the lower bound of $E_x \left( \sum_{j=1}^K \alpha_j^T x \phi'(w_j)[T] x \right)$ when $x$ belongs to Gaussian distribution with zero mean, which is used in proving the lower bound of the Hessian matrix in (81).

**Lemma 12.** Let $f(W; p)$ be the population risk function defined in (17) with $p$ and $W$ satisfying (20). Then, we have

$$\|\nabla^2 f(W[p]; p) - \nabla^2 f(W; p)\|_2 \lesssim \frac{\lambda d^2 + (1 - \lambda)\delta^2}{K} \cdot \frac{\|W[p] - W\|_2}{\sigma_K},$$

(78)

**Lemma 13** (Lemma D.6, [Zhong et al., 2017]). For any $\{w_j\}_{j=1}^K \in \mathbb{R}^d$, let $\alpha \in \mathbb{R}^{dK}$ be the unit vector defined in (19). When the $\phi$ is ReLU function, we have

$$\min_{\|\alpha\|_2 = 1} E_{x \sim \mathcal{N}(0, \sigma^2)} \left( \sum_{j=1}^K \alpha_j^T x \phi'(w_j)[T] x \right)^2 \gtrsim \rho(\sigma),$$

(79)

where $\rho(\sigma)$ is defined in Definition 3.

**Proof of Lemma 7** Let $\lambda_{\max}(W)$ and $\lambda_{\min}(W)$ denote the largest and smallest eigenvalues of $\nabla^2 f(W; p)$ at point $W$, respectively. Then, from Lemma 4 we have

$$\lambda_{\max}(W) \leq \lambda_{\max}(W[p]) + \|\nabla^2 f(W; p) - \nabla^2 f(W[p]; p)\|_2,$$

$$\lambda_{\min}(W) \geq \lambda_{\min}(W[p]) - \|\nabla^2 f(W; p) - \nabla^2 f(W[p]; p)\|_2.$$

(80)

Then, we provide the lower bound of the Hessian matrix of the population function at $W[p]$. For any $\alpha \in \mathbb{R}^{dK}$ defined in (19) with $\|\alpha\|_2 = 1$, we have

$$\min_{\|\alpha\|_2 = 1} \alpha^T \nabla^2 f(W[p]; p) \alpha = \frac{1}{K^2} \min_{\|\alpha\|_2 = 1} \left[ \lambda E_x \left( \sum_{j=1}^K \alpha_j^T x \phi'(w_j[p]T) x \right)^2 + \bar{\lambda} E_x \left( \sum_{j=1}^K \alpha_j^T x \phi'(w_j[p]T - x) \right)^2 \right]$$

$$\geq \frac{1}{K^2} \min_{\|\alpha\|_2 = 1} \lambda E_x \left( \sum_{j=1}^K \alpha_j^T x \phi'(w_j[p]T) x \right)^2 + \min_{\|\alpha\|_2 = 1} \bar{\lambda} E_x \left( \sum_{j=1}^K \alpha_j^T \bar{x} \phi'(w_j[p]T - \bar{x}) \right)^2$$

$$\geq \frac{\lambda \rho(\delta) + \bar{\lambda} \rho(\delta)}{11K^2 \gamma K^2},$$

(81)

where the last inequality comes from Lemma 13.

Next, the upper bound can be bounded as

$$\max_{\|\alpha\|_2 = 1} \alpha^T \nabla^2 f(W[p]; p) \alpha = \frac{1}{K^2} \max_{\|\alpha\|_2 = 1} \left[ \lambda E_x \left( \sum_{j=1}^K \alpha_j^T x \phi'(w_j[p]T) x \right)^2 + \bar{\lambda} E_x \left( \sum_{j=1}^K \alpha_j^T \bar{x} \phi'(w_j[p]T - \bar{x}) \right)^2 \right]$$

$$\leq \frac{1}{K^2} \max_{\|\alpha\|_2 = 1} \lambda E_x \left( \sum_{j=1}^K \alpha_j^T x \phi'(w_j[p]T) x \right)^2 + \max_{\|\alpha\|_2 = 1} \bar{\lambda} E_x \left( \sum_{j=1}^K \alpha_j^T \bar{x} \phi'(w_j[p]T - \bar{x}) \right)^2.$$
For $\mathbb{E}_x \left( \sum_{j=1}^{K} \alpha_j^T x \phi'(w_j^{[p]T} x) \right)^2$, we have
\[
\mathbb{E}_x \left( \sum_{j=1}^{K} \alpha_j^T x \phi'(w_j^{[p]T} x) \right)^2
= \mathbb{E}_x \sum_{j_1=1}^{K} \sum_{j_2=1}^{K} \alpha_{j_1}^T x \phi'(w_{j_1}^{[p]T} x) \alpha_{j_2}^T x \phi'(w_{j_2}^{[p]T} x)
= \sum_{j_1=1}^{K} \sum_{j_2=1}^{K} \mathbb{E}_x \alpha_{j_1}^T x \phi'(w_{j_1}^{[p]T} x) \alpha_{j_2}^T x \phi'(w_{j_2}^{[p]T} x)
\leq \sum_{j_1=1}^{K} \sum_{j_2=1}^{K} \left[ \mathbb{E}_x (\alpha_{j_1}^T x)^4 \mathbb{E}_x (\phi'(w_{j_1}^{[p]T} x))^4 \mathbb{E}_x (\alpha_{j_2}^T x)^4 \mathbb{E}_x (\phi'(w_{j_2}^{[p]T} x))^4 \right]^{1/4}
\leq \sum_{j_1=1}^{K} \sum_{j_2=1}^{K} 3\delta^2 \|\alpha_{j_1}\|_2 \|\alpha_{j_2}\|_2
\leq 6\delta^2 \sum_{j_1=1}^{K} \sum_{j_2=1}^{K} \frac{1}{2} (\|\alpha_{j_1}\|_2^2 + \|\alpha_{j_2}\|_2^2)
= 6K\delta^2.
\]
Therefore, we have
\[
\max_{\|\alpha\|_2=1} \alpha^T \nabla^2 f(W^{[p]}; p) \alpha
\leq \frac{1}{K^2} \max_{\|\alpha\|_2=1} \lambda \mathbb{E}_x \left( \sum_{j=1}^{K} \alpha_j^T x \phi'(w_j^{[p]T} x) \right)^2 + \max_{\|\alpha\|_2=1} \tilde{\lambda} \mathbb{E}_x \left( \sum_{j=1}^{K} \alpha_j^T \tilde{x} \phi'(w_j^{[p]T} \tilde{x}) \right)^2
\leq \frac{6(\delta^2 + \tilde{\delta}^2)}{K}.
\]
Then, given (20), we have
\[
\|W^{(0.0)} - W^{[p]}\|_F = p\|W^{(0.0)} - W\|_F \leq \frac{\sigma_K}{\mu^2 K}.
\]
Combining (85) and Lemma 12 we have
\[
\|\nabla^2 f(W; p) - \nabla^2 f(W^{[p]}; p)\|_2 \leq \frac{\lambda p(\delta) + \tilde{\lambda} p(\tilde{\delta})}{132\kappa^2 \gamma K^2}.
\]
Therefore, (86) and (80) completes the whole proof. □

H.2 Proof of Lemma 2

The task of bounding of the quantity between $\|\nabla \hat{f} - \nabla f\|_2$ is dividing into bounding $I_1$, $I_2$, $I_3$ and $I_4$ as shown in (89). $I_1$ and $I_3$ represent the deviation of the mean of several random variables to their expectation, which can be bounded through concentration inequality, i.e., Chernoff bound. $I_2$ and $I_4$ come from the inconsistency of the output label $y$ and pseudo label $\tilde{y}$ in the empirical risk function in (1) and population risk function in (17). The major challenge lies in characterizing the upper bound of $I_2$ and $I_4$ as the linear function of $W - W^{[p]}$ and $W^{[p]} - W^*$, which is summarized in (86).
Proof of Lemma 2. From (1), we know that

\[
\frac{\partial f}{\partial w_k}(W) = \frac{\lambda}{N} \sum_{n=1}^{N} \left( \sum_{j=1}^{K} \phi(w_j^T x_n - y_n) x_n + \frac{1 - \lambda}{M} \sum_{m=1}^{M} \left( \sum_{j=1}^{K} \phi(w_j^T \bar{x}_m - \bar{y}_n) \bar{x}_m \right) \right)
\]

\[
= \frac{\lambda}{K^2 N} \sum_{n=1}^{N} \sum_{j=1}^{K} \left( \phi(w_j^T x_n - \phi(w_j^T x_n)) x_n \right)
\]

\[
+ \frac{1 - \lambda}{K^2 M} \sum_{m=1}^{M} \left( \sum_{j=1}^{K} \phi(w_j^T \bar{x}_m - \phi(w_j^T \bar{x}_m)) \bar{x}_m \right)
\]

From (32), we know that

\[
\frac{\partial f}{\partial w_k}(W) = \lambda \frac{1}{K^2} \sum_{j=1}^{K} \left( \phi(w_j^T x) - \phi(w_j^T x) \right) x + \frac{1 - \lambda}{K^2} \sum_{j=1}^{K} \left( \phi(w_j^T \bar{x}) - \phi(w_j^T \bar{x}) \right) \bar{x}
\]

Then, from (17), we have

\[
\frac{\partial f}{\partial w_k}(W) = \frac{\partial f}{\partial w_k}(W; p)
\]

\[
= \lambda \frac{1}{K^2 N} \sum_{n=1}^{N} \left( \phi(w_j^T x_n) - \phi(w_j^T x_n) \right) x_n \sum_{j=1}^{K} \left( \phi(w_j^T x_n) - \phi(w_j^T x_n) \right) \bar{x}_m
\]

\[
+ \frac{1 - \lambda}{K^2 M} \sum_{m=1}^{M} \left( \phi(w_j^T \bar{x}_m) - \phi(w_j^T \bar{x}_m) \right) \bar{x}_m \sum_{j=1}^{K} \left( \phi(w_j^T \bar{x}_m) - \phi(w_j^T \bar{x}_m) \right) \bar{x}
\]

\[
= \lambda \frac{1}{K^2} \sum_{j=1}^{K} \left( \phi(w_j^T x) - \phi(w_j^T x) \right) x + \frac{1 - \lambda}{K^2} \sum_{j=1}^{K} \left( \phi(w_j^T \bar{x}) - \phi(w_j^T \bar{x}) \right) \bar{x}
\]

\[
:= I_1 + I_2 + I_3 + I_4.
\]

For any \(\alpha_j \in \mathbb{R}^d\) with \(||\alpha_j||_2 \leq 1\), we define a random variable \(Z(j) = (\phi(w_j^T x) - \phi(w_j^T x))\alpha_j^T x\) and \(Z_n(j) = (\phi(w_j^T x_n) - \phi(w_j^T x_n))\alpha_j^T x_n\) as the realization of \(Z(j)\) for \(n = 1, 2, \ldots, N\). Then, for any \(p \in \mathbb{N}^+\), we have

\[
(E|Z|^p)^{1/p} = \left( E(\phi(w_j^T x) - \phi(w_j^T x))^{p} \cdot |\alpha_j^T x|^{p} \right)^{1/p}
\]

\[
\leq C \cdot \delta^2 ||w_j - w_j^*||_2 \cdot p,
\]

where \(C\) is a positive constant and the last inequality holds since \(x \sim \mathcal{N}(0, \delta^2)\). From Definition 2, we know that \(Z\) belongs to sub-exponential distribution with \(||Z||_{\psi_1} \leq \delta^2 ||w_j - w_j^*||_2\). Therefore, by Chernoff inequality, we have

\[
P \left( \left| \frac{1}{N} \sum_{n=1}^{N} Z_n(j) - EZ(j) \right| < t \right) \leq 1 - e^{-\frac{C(\delta^2 ||w_j - w_j^*||_2)^2 \cdot N \cdot s^2}{e^{Nst}}}
\]
for some positive constant $C$ and any $s \in \mathbb{R}$.

Let $t = \delta^2 \|w_j - w_j^*\|_2 \sqrt{\frac{d \log q}{N}}$ and $s = \frac{2}{C \delta^2 \|w_j - w_j^*\|_2} \cdot t$ for some large constant $q > 0$, we have

$$\left| \frac{1}{N} \sum_{n=1}^{N} Z_n(j) - \mathbb{E}Z(j) \right| \lesssim \delta^2 \|w_j - w_j^*\|_2 \cdot \sqrt{\frac{d \log q}{N}} \quad \text{(92)}$$

with probability at least $1 - q^{-d}$. From Lemma 7, we have

$$\left\| \frac{1}{M} \sum_{m=1}^{M} (\phi(w_j^T x_m) - \phi(w_j^{[p]} T x_m)) x_n - \mathbb{E}x \left( \phi(w_j^T x) - \phi(w_j^{[p]} T x) \right) x \right\|_2 \lesssim 2 \delta^2 \|w_j - w_j^*\|_2 \cdot \sqrt{\frac{d \log q}{M}} \quad \text{(93)}$$

with probability at least $1 - (q/5)^{-d}$. Since $q$ is a large constant, we release the probability as $1 - q^{-d}$ for simplification. Similar to $Z$, we have

$$\left\| \frac{1}{M} \sum_{m=1}^{M} (\phi(w_j^T x_m) - \phi(w_j^{[p]} T x_m)) x_n - \mathbb{E}x \left( \phi(w_j^T x) - \phi(w_j^{[p]} T x) \right) x \right\|_2 \lesssim 2 \delta^2 \|w_j - w_j^*\|_2 \cdot \sqrt{\frac{d \log q}{M}} \quad \text{(94)}$$

with probability at least $1 - q^{-d}$.

**Figure 13:** The subspace spanned by $w_j^*$ and $w_j^{[p]}$

For term $\mathbb{E}_x \left[ (\phi(w_j^T x) - \phi(w_j^{[p]} T x)) x \right]$, let us define the angle between $w_j^*$ and $w_j^{[p]}$ as $\theta_1$. Figure 13 shows the subspace spanned by the vector $w_j^*$ and $\bar{w}_j$. We divide the subspace by 4 pieces, where the gray region denotes area I, and the blue area denotes area II. Areas III and IV are the symmetries of II and I from the origin, respectively. Hence, we have

$$\begin{align*}
\mathbb{E}_x \left[ (\phi(w_j^T x) - \phi(w_j^{[p]} T x)) x \right] &= \mathbb{E}_{x \in \text{area I}} \left[ (\phi(w_j^T x) - \phi(w_j^{[p]} T x)) x \right] + \mathbb{E}_{x \in \text{area II}} \left[ (\phi(w_j^T x) - \phi(w_j^{[p]} T x)) x \right] \\
&+ \mathbb{E}_{x \in \text{area III}} \left[ (\phi(w_j^T x) - \phi(w_j^{[p]} T x)) x \right] + \mathbb{E}_{x \in \text{area IV}} \left[ (\phi(w_j^T x) - \phi(w_j^{[p]} T x)) x \right] \\
&= \mathbb{E}_{x \in \text{area I}} \left[ (\phi(w_j^T x) - \phi(w_j^{[p]} T x)) x \right] + \mathbb{E}_{x \in \text{area II}} \left[ w_j^{[p]} T x \bar{x} \right] - \mathbb{E}_{x \in \text{area III}} \left[ w_j^{[p]} T x \bar{x} \right] \\
&= \frac{1}{2} \mathbb{E}_x \left[ (w_j^* - w_j^{[p]} T x) x \right] + \mathbb{E}_{x \in \text{area II}} \left[ (w_j^* - w_j^{[p]} T x) x \right] \\
&= \frac{1}{2} \mathbb{E}_x \left[ (w_j^* - w_j^{[p]} T x) x \right]
\end{align*} \quad \text{(95)}$$

29
Therefore, we have
\[
\begin{align*}
\frac{\lambda}{K^2} & \mathbb{E}_x \left[ (\phi(w_j^T x) - \phi(w_j^{[p]}T x)) x \right] + \frac{1 - \lambda}{K^2} \mathbb{E}_x \left[ (\phi(\tilde{w}_j^T x) - \phi(w_j^{[p]}T x)) x \right] \\
= & \frac{\lambda}{2K^2} \mathbb{E}_x \left[ (w_j - w_j^{[p]})^T x x \right] + \frac{1 - \lambda}{2K^2} \mathbb{E}_x \left[ (\tilde{w}_j - w_j^{[p]})^T x x \right] \\
= & \frac{\|\lambda \delta^2 \cdot (\tilde{w}_j - w_j^{[p]}) + (1 - \lambda) \delta^2 \cdot (w_j - w_j^{[p]}) \|^2}{2K^2}.
\end{align*}
\] (96)

From (93), (94) and (96), we have
\[
\begin{align*}
\left\| \frac{\partial \hat{f}}{\partial w_k}(W; p) - \frac{\partial \hat{f}}{\partial w_k}(W) \right\|_2 \\
\leq & \frac{\lambda}{K^2} \sum_{j=1}^K \sum_{n=1}^N \left\| \left( \frac{\partial \phi}{\partial w_j^T} x_n - \frac{\partial \phi}{\partial w_j^{[p]}T x_n} \right) x_n - \mathbb{E}_x \left( \frac{\partial \phi}{\partial w_j^T} x - \frac{\partial \phi}{\partial w_j^{[p]}T x} \right) x \right\|_2 \\
& + \frac{1 - \lambda}{K^2} \sum_{j=1}^K \left\| \left( \frac{\partial \phi}{\partial w_j^T \tilde{x}_m} - \frac{\partial \phi}{\partial w_j^{[p]}T \tilde{x}_m} \right) \tilde{x}_m - \mathbb{E}_x \left( \frac{\partial \phi}{\partial w_j^T \tilde{x}} - \frac{\partial \phi}{\partial w_j^{[p]}T \tilde{x}} \right) \tilde{x} \right\|_2 \\
& + \sum_{j=1}^K \left\| \frac{\lambda}{K^2} \mathbb{E}_x \left[ (\phi(w_j^* T x) - \phi(w_j^{[p]}T x)) x \right] + \frac{1 - \lambda}{K^2} \mathbb{E}_x \left[ (\phi(\tilde{w}_j^T x) - \phi(w_j^{[p]}T x)) \tilde{x} \right] \right\|_2 \\
\leq & \frac{\lambda}{K^2} \delta^2 \sqrt{\frac{d \log q}{N}} \cdot \sum_{j=1}^K \|w_j - w_j^*\|_2 + \frac{1 - \lambda}{K^2} \cdot \delta^2 \sqrt{\frac{d \log q}{M}} \cdot \sum_{j=1}^K \|w_j - \tilde{w}_j\|_2 \\
& + \frac{1}{2K^2} \cdot \sum_{j=1}^K \|\lambda \delta^2 \cdot (\tilde{w}_j - w_j^{[p]}) + (1 - \lambda) \delta^2 \cdot (w_j - w_j^{[p]})\|_2 \\
\leq & \frac{\lambda}{K^{3/2}} \delta^2 \sqrt{\frac{d \log q}{N}} \cdot \|W - W^*\|_2 + \frac{1 - \lambda}{K^{3/2}} \cdot \delta^2 \sqrt{\frac{d \log q}{M}} \cdot \|W - \tilde{W}\|_2 \\
& + \frac{1}{2K^{1/2}} \|\lambda \delta^2 \cdot (\tilde{W} - W^{[p]}) + (1 - \lambda) \delta^2 \cdot (W^* - W^{[p]})\|_2
\end{align*}
\] (97)

with probability at least \(1 - q^{-d}\).

In conclusion, let \(\alpha \in \mathbb{R}^{Kd}\) and \(\alpha_j \in \mathbb{R}^d\) with \(\alpha = [\alpha_1^T, \alpha_2^T, \cdots, \alpha_K^T]^T\), we have
\[
\left\| \nabla f(W) - \nabla \hat{f}(W) \right\|_2 = \left| \alpha^T (\nabla f(W) - \nabla \hat{f}(W)) \right| \\
\leq & \sum_{k=1}^K \left| \alpha_k^T \left( \frac{\partial \hat{f}}{\partial w_k}(W) - \frac{\partial f}{\partial w_k}(W) \right) \right| \\
\lesssim & \sum_{k=1}^K \left\| \frac{\partial \hat{f}}{\partial w_k}(W) - \frac{\partial f}{\partial w_k}(W) \right\|_2 \cdot \|\alpha_k\|_2 \\
\leq & \frac{\lambda}{K} \delta^2 \sqrt{\frac{d \log q}{N}} \cdot \|W - W^*\|_2 + \frac{1 - \lambda}{K} \cdot \delta^2 \sqrt{\frac{d \log q}{M}} \cdot \|W - \tilde{W}\|_2 \\
& + \frac{1}{2K} \|\lambda \delta^2 \cdot (\tilde{W} - W^{[p]}) + (1 - \lambda) \delta^2 \cdot (W^* - W^{[p]})\|_2
\] (98)

with probability at least \(1 - q^{-d}\). \(\square\)

**H.3 Proof of Lemma 12**

The distance of the second order derivatives of the population risk function \(f(\cdot; p)\) at point \(W\) and \(W^{[p]}\) can be converted into bounding \(P_1, P_2, P_3\) and \(P_4\), which are defined in (101). The major
idea in proving \( P_1 \) is to connect the error bound to the angle between \( W \) and \( W^{[p]} \). Similar ideas apply in bounding the other three items as well.

**Proof of Lemma[72]** From [17], we have

\[
\frac{\partial^2 f}{\partial w_{j1} \partial w_{j2}} (W^{[p]}; p) = \frac{\lambda}{K^2} E_x \phi'((w^{[p]}_j x) \phi'(w^{[p]}_j x) x x^T + \frac{1}{K^2} \frac{1}{E_x} \phi'((w^{[p]}_j x) \phi'(w^{[p]}_j x) x x^T, \\
\text{and } \frac{\partial^2 f}{\partial w_{j1} \partial w_{j2}} (W; p) = \frac{\lambda}{K^2} E_x \phi'((w^{[p]}_j x) \phi'(w^{[p]}_j x) x x^T + \frac{1}{K^2} \frac{1}{E_x} \phi'((w^{[p]}_j x) \phi'(w^{[p]}_j x) x x^T,}
\]

where \( w^{[p]}_j \) is the \( j \)-th column of \( W^{[p]} \). Then, we have

\[
\frac{\partial^2 f}{\partial w_{j1} \partial w_{j2}} (W) - \frac{\partial^2 f}{\partial w_{j1} \partial w_{j2}} (W) = \lambda \frac{E_x}{K^2} \left[ \phi'((w^{[p]}_j x) \phi'(w^{[p]}_j x) x x^T + \frac{1}{K^2} \frac{1}{E_x} \phi'((w^{[p]}_j x) \phi'(w^{[p]}_j x) x x^T.
\]

For any \( a \in \mathbb{R}^d \) with \( \|a\|_2 = 1 \), we have

\[
\alpha^T P_1 \alpha = E_x \phi'((w^{[p]}_j x) \phi'(w^{[p]}_j x) - \phi'((w^{[p]}_j x) (a^T x)^2.
\]

where \( a \in \mathbb{R}^d \). Let \( I = \phi'((w^{[p]}_j x) \phi'(w^{[p]}_j x) - \phi'((w^{[p]}_j x) (a^T x)^2 \). It is easy to verify there exists a group of orthonormal vectors such that \( B = \{ a, b, c, a_1, \cdots, a_d \} \) with \( a, b, c \) spans a subspace that contains \( a, w_{j1}, \text{and } w_{j2} \). Then, for any \( x \), we have a unique \( z = [z_1, z_2, \cdots, z_d] \) such that

\[
x = z_1 a + z_2 b + z_3 c + \cdots + z_d a_d.
\]

Also, since \( x \sim N(0, \delta^2 I_d) \), we have \( z \sim N(0, \delta^2 I_d) \). Then, we have

\[
I = E_{z_1, z_2, z_3} |\phi'((w^{[p]}_j x) - \phi'((w^{[p]}_j x)| a^T x|^2
\]

where \( x = z_1 a + z_2 b + z_3 c \) and \( f_z(z_1, z_2, z_3) \) is probability density function of \( (z_1, z_2, z_3) \). Next, we consider spherical coordinates with \( z_1 = R \cos \phi_1, z_2 = R \sin \phi_1 \sin \phi_2, z_3 = R \sin \phi_1 \cos \phi_2 \). Hence,

\[
I = \int |\phi'((w^{[p]}_j x) - \phi'((w^{[p]}_j x)| R \cos \phi_1 |^2 \cdot f_z(z_1, z_2, z_3) d_z_1 d_z_2 d_z_3.
\]

It is easy to verify that \( \phi'((w^{[p]}_j x) \) only depends on the direction of \( x \) and

\[
f_z(R, \phi_1, \phi_2) = \frac{1}{(2\pi \delta^2)^{\frac{d}{2}}} e^{-\frac{R^2}{2\delta^2}} = \frac{1}{(2\pi \delta^2)^{\frac{d}{2}}} e^{-\frac{R^2}{2\delta^2}}
\]

31
only depends on $R$. Then, we have
\[
I(i_2, j_2) = \int \phi'(w_{j_2}^T(x/R)) - \phi'(w_{j_2}^{[p]T}(x/R)) \cdot |R \cos \phi_1|^2 \cdot f_Z(R) R^2 \sin \phi_1 \, dR \, d\phi_1 d\phi_2
\]
\[
= \int_0^\infty R^2 f_Z(R) dR \int_0^\pi \int_0^{2\pi} |\cos \phi_1|^2 \cdot \sin \phi_1 \cdot |\phi'(w_{j_2}^T(x/R)) - \phi'(w_{j_2}^{[p]T}(x/R))| \, d\phi_1 d\phi_2
\]
\[
\leq 3\delta^2 \cdot \int_0^\infty R^2 f_Z(R) dR \int_0^\pi \int_0^{2\pi} \sin \phi_1 \cdot |\phi'(w_{j_2}^T(x/R)) - \phi'(w_{j_2}^{[p]T}(x/R))| \, d\phi_1 d\phi_2
\]
\[
= 3\delta^2 \cdot \mathbb{E}_{z_1, z_2, z_3} |\phi'(w_{j_2}^T x) - \phi'(w_{j_2}^{[p]T} x)|
\]
\[
\leq 3\delta^2 \cdot \mathbb{E}_{x} |\phi'(w_{j_2}^T x) - \phi'(w_{j_2}^{[p]T} x)|,
\]
(104)

where the inequality (a) is derived from the fact that $|\cos \phi_1| \leq 1$ and
\[
\int_0^\infty R^2 f_Z(R) dR = \int_0^\infty \frac{1}{(2\pi\delta^2)^2} e^{-\frac{R^2}{2\delta^2}} dR = \int_0^\infty \frac{1}{(2\pi\delta^2)^2} d(e^{-\frac{R^2}{2\delta^2}})
\]
\[
= \int_0^\infty e^{-\frac{R^2}{2\delta^2}} \frac{R^2}{(2\pi\delta^2)^2} dR = 3\delta^2 \int_0^\infty R^2 e^{-\frac{R^2}{2\delta^2}} dR.
\]
(105)

Define a set $A_1 = \{x | (w_{j_2}^{[p]T} x)(w_{j_2}^T x) < 0\}$. If $x \in A_1$, then $w_{j_2}^{[p]T} x$ and $w_{j_2}^T x$ have different signs, which means the value of $\phi'(w_{j_2}^T x)$ and $\phi'(w_{j_2}^{[p]T} x)$ are different. This is equivalent to say that
\[
|\phi'(w_{j_2}^T x) - \phi'(w_{j_2}^{[p]T} x)| = \begin{cases} 1, & \text{if } x \in A_1 \\ 0, & \text{if } x \in A_1^c \end{cases}
\]
(106)

Moreover, if $x \in A_1$, then we have
\[
|w_{j_2}^{[p]T} x| \leq |w_{j_2}^T x - w_{j_2}^{[p]T} x| \leq \|w_{j_2}^{[p]} - w_{j_2}\|_2 \cdot \|x\|_2.
\]
(107)

Let us define a set $A_2$ such that
\[
A_2 = \{x \mid \frac{|w_{j_2}^{[p]} x|}{\|w_{j_2}^{[p]}\|_2 \cdot \|x\|_2} \leq \frac{\|w_{j_2}^T - w_{j_2}^{[p]}\|_2}{\|w_{j_2}^{[p]}\|_2} \} = \left\{ \theta_{x, w_{j_2}^{[p]}} \mid \cos \theta_{x, w_{j_2}^{[p]}} \leq \frac{\|w_{j_2}^{[p]} - w_{j_2}\|_2}{\|w_{j_2}^{[p]}\|_2} \right\}.
\]
(108)

Hence, we have that
\[
\mathbb{E}_x |\phi'(w_{j_2}^T x) - \phi'(w_{j_2}^{[p]T} x)|^2 = \mathbb{E}_x |\phi'(w_{j_2}^T x) - \phi'(w_{j_2}^{[p]T} x)|
\]
\[
= \text{Prob}(x \in A_1)
\]
\[
\leq \text{Prob}(x \in A_2).
\]
(109)

Since $x \sim \mathcal{N}(0, \delta^2 \|a\|_2^2 I)$, $\theta_{x, w_{j_2}^{[p]}}$ belongs to the uniform distribution on $[-\pi, \pi]$, we have
\[
\text{Prob}(x \in A_2) = \frac{\pi - \arccos \frac{\|w_{j_2}^{[p]} - w_{j_2}\|_2}{\|w_{j_2}^{[p]}\|_2}}{\pi} \leq \frac{1}{\pi} \tan(\pi - \arccos \frac{\|w_{j_2}^{[p]} - w_{j_2}\|_2}{\|w_{j_2}^{[p]}\|_2})
\]
\[
= \frac{1}{\pi} \cot(\arccos \frac{\|w_{j_2}^{[p]} - w_{j_2}\|_2}{\|w_{j_2}^{[p]}\|_2}) \leq \frac{2}{\pi} \frac{\|w_{j_2}^{[p]} - w_{j_2}\|_2}{\|w_{j_2}^{[p]}\|_2}.
\]
(110)
Hence, \([10]\) and \([10]\) suggest that
\[
I \leq \frac{\lambda}{\sigma_K} \frac{\|w_j - w_i\|_2}{\|a\|_2}.
\] (111)

The same bound that shown in \([111]\) holds for \(P_2\) as well. \(P_3\) and \(P_4\) satisfy \([111]\) except for changing \(\delta^2\) to \(\tilde{\delta}^2\).

Therefore, we have
\[
\|\nabla^2 f(W[p]; p) - \nabla^2 f(W; p)\|_2
\]
\[
= \max_{\|\alpha\|_2 \leq 1} \left| \alpha^T (\nabla^2 f(W[p]; p) - \nabla^2 f(W; p)\alpha) \right|
\]
\[
\leq \sum_{j=1}^{K} \sum_{j=1}^{K} \left| \alpha_{j1}^T \left( \frac{\partial^2 f}{\partial w_{j1} \partial w_{j2}} (W[p]; p) - \frac{\partial^2 f}{\partial w_{j1} \partial w_{j2}} (W; p) \right) \alpha_{j2} \right|
\]
\[
\leq \frac{1}{K^2} \sum_{j=1}^{K} \sum_{j=1}^{K} \left( \lambda \|P_1 + P_2\|_2 + (1 - \lambda) \|P_3 + P_4\|_2 \right) \|\alpha_{j1}\|_2 \|\alpha_{j2}\|_2
\]
\[
\leq \frac{1}{K^2} \sum_{j=1}^{K} \sum_{j=1}^{K} \lambda (\|P_1 + P_2\|_2 + (1 - \lambda) \|P_3 + P_4\|_2) \frac{\|w_j - w_i\|_2}{\sigma_K} \|\alpha_{j1}\|_2 \|\alpha_{j2}\|_2
\]
\[
\leq \frac{4}{K} \left( \lambda \|P_1 + P_2\|_2 + (1 - \lambda) \|P_3 + P_4\|_2 \right) \frac{\|W[p] - W\|_2}{\sigma_K},
\] (112)

where \(\alpha \in \mathbb{R}^{Kd}\) and \(\alpha_j \in \mathbb{R}^d\) with \(\alpha = [\alpha_1^T, \alpha_2^T, \cdots, \alpha_K^T]^T\).

I INITIALIZATION VIA TENSOR METHOD

In this section, we briefly summarize the tensor initialization in \([\text{Zhong et al.} \ 2017]\) by studying the target function class as
\[
y = \frac{1}{K} \sum_{j=1}^{K} v_j^* \phi(w_j^T x),
\] (113)

where \(v_j^* \in R\). Note that for ReLU function, we have \(v_j^* \phi(w_j^T x) = \text{sign}(v_j^*) \phi(|w_j| w_j^T x)\). Without loss of generalization, we can assume \(v_j^* \in \{+1, -1\}\). Additionally, it is clear that the function studied in \([8]\) is the special case of \([113]\) when \(v_j^* = 1\) for all \(j\). In addition, Theorem 5.6 in \([\text{Zhong et al.} \ 2017]\) show that the sign of \(v_j^*\) can be directly recovered using tensor initialization, which indicates the equivalence of \([8]\) and \([113]\) when using tensor initialization.

We first define some high order momenta in the following way:
\[
M_1 = \mathbb{E}_x \{y x\} \in \mathbb{R}^d,
\] (114)
\[
M_2 = \mathbb{E}_x \left[ y (x \otimes x - \delta^2 I) \right] \in \mathbb{R}^{d \times d},
\] (115)
\[
M_3 = \mathbb{E}_x \left[ y (x \otimes^3 - x \otimes^2 \delta^2 I) \right] \in \mathbb{R}^{d \times d \times d},
\] (116)

where \(\mathbb{E}_x\) is the expectation over \(x\) and \(z \otimes^3 := z \otimes z \otimes z\). The operator \(\otimes\) is defined as
\[
v \otimes Z = \sum_{i=1}^{d_z} (v \otimes z_i \otimes z_i + z_i \otimes v \otimes z_i + z_i \otimes z_i \otimes v),
\] (117)

for any vector \(v \in \mathbb{R}^{d_z}\) and \(Z \in \mathbb{R}^{d_z \times d_z}\).
Following the same calculation formulas in the Claim 5.2 (Zhong et al. [2017]), there exist some known constants $\psi_i, i = 1, 2, 3$, such that

$$M_1 = \sum_{j=1}^{K} \psi_1 \cdot \|w_j^*\|_2 \cdot \vec{w}_j^*,$$

$$M_2 = \sum_{j=1}^{K} \psi_2 \cdot \|w_j^*\|_2 \cdot \vec{w}_j^* \vec{w}_j^T,$$

$$M_3 = \sum_{j=1}^{K} \psi_3 \cdot \|w_j^*\|_2 \cdot \vec{w}_j^* \otimes^3,$$

(118) (119) (120)

where $\vec{w}_j^* = w_j^*/\|w_j^*\|_2$ in [114]-[116] is the normalization of $w_j^*$. Therefore, we can see that the information of $\{w_j^*\}_{j=1}^{K}$ is separated as the direction of $w_j$ and the magnitude of $w_j$ in $M_1$, $M_2$ and $M_3$.

$M_1$, $M_2$ and $M_3$ can be estimated through the samples $\{(x_n, y_n)\}_{n=1}^{N}$, and let $\tilde{M}_1$, $\tilde{M}_2$, $\tilde{M}_3$ denote the corresponding estimates. First, we will decompose the rank-$K$ tensor $M_3$ and obtain the $\{\vec{w}_j^*\}_{j=1}^{K}$. By applying the tensor decomposition method (Kuleshov et al. [2015]) to $\tilde{M}_3$, the outputs, denoted by $\vec{w}_j^*$, are the estimations of $\{s_j \vec{w}_j^*\}_{j=1}^{K}$, where $s_j$ is an unknown sign. Second, we will estimate $s_j$, $v_j^*$ and $\|w_j^*\|_2$ through $M_1$ and $M_2$. Note that $M_2$ does not contain the information of $s_j$ because $s_j^2$ is always 1. Then, through solving the following two optimization problem:

$$\hat{\alpha}_1 = \arg \min_{\alpha_1 \in \mathbb{R}^K} \left| \tilde{M}_1 - \sum_{j=1}^{K} \psi_1 \alpha_{1,j} \vec{w}_j^* \right|,$$

$$\hat{\alpha}_2 = \arg \min_{\alpha_2 \in \mathbb{R}^K} \left| \tilde{M}_2 - \sum_{j=1}^{K} \psi_2 \alpha_{2,j} \vec{w}_j^* \vec{w}_j^* \right|,$$

(121)

The estimation of $s_j$ can be given as

$$\hat{s}_j = \text{sign}(\hat{\alpha}_{1,j}/\hat{\alpha}_{2,j}).$$

Also, we know that $|\hat{\alpha}_{1,j}|$ is the estimation of $\|w_j^*\|$ and

$$\hat{v}_j = \text{sign}(\hat{\alpha}_{1,j}/s_j) = \text{sign}(\hat{\alpha}_{2,j}).$$

Thus, $W(0)$ is given as

$$[\text{sign}(\hat{\alpha}_{2,1})\hat{\alpha}_{1,1}\vec{w}_1^*, \ldots, \text{sign}(\hat{\alpha}_{2,K})\hat{\alpha}_{1,K}\vec{w}_K^*].$$

**Subroutine 1 Tensor Initialization Method**

1: **Input:** labeled data $\mathcal{D} = \{(x, y)\}_{n=1}^{N}$
2: Partition $\mathcal{D}$ into three disjoint subsets $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_3$;
3: Calculate $\tilde{M}_1$, $\tilde{M}_2$ following [114], [115] using $\mathcal{D}_1$, $\mathcal{D}_2$, respectively;
4: Obtain the estimate subspace $\hat{V}$ of $M_2$;
5: Calculate $\tilde{M}_3(\hat{V}, \hat{V}, \hat{V})$ through $\mathcal{D}_3$;
6: Obtain $\{\hat{s}_j\}_{j=1}^{K}$ via tensor decomposition method (Kuleshov et al. [2015]) on $\hat{M}_3(\hat{V}, \hat{V}, \hat{V})$;
7: Obtain $\hat{\alpha}_1$, $\hat{\alpha}_2$ by solving optimization problem (121);
8: **Return:** $w_j^{(0)} = \text{sign}(\hat{\alpha}_{2,j})\hat{\alpha}_{1,j}\hat{V}\vec{u}_j$ and $v_j^{(0)} = \text{sign}(\hat{\alpha}_{2,j})$, $j = 1, ..., K$.

To reduce the computational complexity of tensor decomposition, one can project $\hat{M}_3$ to a lower-dimensional tensor (Zhong et al. [2017]). The idea is to first estimate the subspace spanned by $\{w_j^*\}_{j=1}^{K}$, and let $\hat{V}$ denote the estimated subspace. Moreover, we have

$$M_3(\hat{V}, \hat{V}, \hat{V}) = \mathbb{E}_x \left[ y((\hat{V}^T x)^{\otimes 3} - (\hat{V}^T x)^{\otimes 2} \mathbb{E}_y (\hat{V}^T x)(\hat{V}^T x)^T) \right] \in \mathbb{R}^{K \times K \times K},$$

(122)
Then, one can decompose the estimate \( \hat{M}(\mathbf{V}, \mathbf{V}, \mathbf{V}) \) to obtain unit vectors \( \{ \hat{s}_j \}_{j=1}^{K} \in \mathbb{R}^K \). Since \( \mathbf{w}^* \) lies in the subspace \( \mathbf{V} \), we have \( \mathbf{VV}^T \mathbf{w}_j^* = \mathbf{w}_j^* \). Then, \( \mathbf{V} \hat{s}_j \) is an estimate of \( \mathbf{w}_j^* \). The initialization process is summarized in Subroutine

### J  Classification Problems

The framework in this paper is extendable to binary classification problem. For binary classification problem, the output \( y \) given input \( \mathbf{x} \) is defined as

\[
\text{Prob}\{y = 1\} = g(\mathbf{W}^*; \mathbf{x})
\]

with some ground truth parameter \( \mathbf{W}^* \). To guarantee the output is within \([0, 1]\), the activation function is often used as sigmoid. For classification, the loss function is cross-entropy, and the objective function over labeled data \( \mathcal{D} \) is defined as

\[
f_{\mathcal{D}}(\mathbf{W}) = \frac{1}{N} \sum_{(\mathbf{x}_n, y_n) \in \mathcal{D}} -y_n \log g(\mathbf{W}; \mathbf{x}_n) - (1 - y_n) \log(1 - g(\mathbf{W}; \mathbf{x}_n)).
\]

The expectation of objective function can be written as

\[
E_{\mathcal{D}}f_{\mathcal{D}}(\mathbf{W}) = E_{(\mathbf{x}, y)} - y \log(g(\mathbf{W}; \mathbf{x}_n)) - (1 - y) \log(1 - g(\mathbf{W}; \mathbf{x}))
\]

\[
= E_{\mathbf{x}} [ -y \log(g(\mathbf{W}^*; \mathbf{x}_n)) - (1 - y) \log(1 - g(\mathbf{W}^*; \mathbf{x})) ]
\]

Please note that (125) is exactly the same as (32) with \( \lambda = 1 \) when the loss function is squared loss. For cross entropy loss function, the second order derivative of (125) is calculated as

\[
\frac{\partial f_{\mathcal{D}}(\mathbf{W})}{\partial w_j \partial w_k} = \frac{1}{N} \left[ \frac{y_n}{g^2(\mathbf{W}; \mathbf{x})} + \frac{(1 - y_n)}{(1 - g(\mathbf{W}; \mathbf{x}))^2} \right] \cdot \phi'(w_j^T \mathbf{x}) \phi'(w_k^T \mathbf{x}) \mathbf{x} \mathbf{x}^T.
\]

when \( j \neq k \). Refer to (88) in (Fu et al., 2020) or (132) in (Zhang et al., 2020b), we have

\[
\left\| \frac{y_n (\phi'(w_j^T \mathbf{x}) \phi'(w_k^T \mathbf{x}))}{g^2(\mathbf{W}; \mathbf{x})} \right\|_2 \leq \left\| \frac{\phi'(w_j^T \mathbf{x}) \phi'(w_k^T \mathbf{x})}{g^2(\mathbf{W}; \mathbf{x})} \right\|_2 \leq K^2.
\]

Following similar steps in (90), from Definition 2, we know that \( \alpha_j^T \frac{\partial f_{\mathcal{D}}(\mathbf{W})}{\partial w_j} \alpha_k \) belongs to the sub-exponential distribution. Therefore, similar results for objective function with cross-entropy loss can be established as well. One can check (Fu et al., 2020) or (Zhang et al., 2020b) for details.

### K  One-Hidden Layer Neural Network with Top Layer Weights

For a general one-hidden layer neural network, the output of the neural network is defined as

\[
g(\mathbf{W}, \mathbf{v}; \mathbf{x}) = \frac{1}{K} \sum_{j=1}^{K} v_j \phi(w_j^T \mathbf{x}),
\]

where \( \mathbf{v} = [v_1, v_2, \cdots, v_K] \in \mathbb{R}^K \). Then, the target function can be defined as

\[
y = g(\mathbf{W}^*, \mathbf{v}^*; \mathbf{x}) = \frac{1}{K} \sum_{j=1}^{K} v_j^* \phi(w_j^T \mathbf{x})
\]

for some unknown weights \( \mathbf{W}^* \) and \( \mathbf{v}^* \).

In the following paragraphs, we will provide a short description for the equivalence of (129) and (2) in theoretical analysis. Note that for ReLU functions, we have \( v_j \phi(w_j^T \mathbf{x}) = \text{sign}(v_j) \phi(|v_j| w_j^T \mathbf{x}) \).
Without loss of generalization, we can assume $v_j, v_j^* \in \{+1, -1\}$ for all $j \in [K]$. From Appendix A, we know that the sign of $v_j^*$ can exactly estimated through tensor initialization. There, we can focus on analysis the neural network in the form as

$$g(W; x) = \frac{1}{K} \sum_{j=1}^{K} v_j^* \phi(w_j^T x).$$  \hspace{1cm} (130)

Considering the objective function in $\Box$ and population risk function in (17), we have

$$\| \frac{\partial f}{\partial w_k}(W) - \frac{\partial f}{\partial w_k}(W; \tilde{x}) \|_2 = \left\| \frac{\lambda}{K^2 N} \sum_{j=1}^{K} v_j^* \left[ \sum_{n=1}^{N} \left( \phi(w_j^T x_n) - \phi(w_j^* T x_n) \right) x_n - E_x \left( \phi(w_j^T x) - \phi(w_j^* T x) \right) x \right] 
+ \frac{1 - \lambda}{K^2 M} \sum_{j=1}^{K} \left[ \sum_{m=1}^{M} \left( \phi(w_j^T \tilde{x}_m) - \phi(w_j^* T \tilde{x}_m) \right) \tilde{x}_m - E_{\tilde{x}} \left( \phi(w_j^T \tilde{x}) - \phi(w_j^* T \tilde{x}) \right) \tilde{x} \right] \right\|_2
\leq \sum_{j=1}^{K} |v_j^*| \left\| \frac{\lambda}{K^2 N} \left[ \sum_{n=1}^{N} \left( \phi(w_j^T x_n) - \phi(w_j^* T x_n) \right) x_n - E_x \left( \phi(w_j^T x) - \phi(w_j^* T x) \right) x \right] 
+ \frac{1 - \lambda}{K^2 M} \left[ \sum_{m=1}^{M} \left( \phi(w_j^T \tilde{x}_m) - \phi(w_j^* T \tilde{x}_m) \right) \tilde{x}_m - E_{\tilde{x}} \left( \phi(w_j^T \tilde{x}) - \phi(w_j^* T \tilde{x}) \right) \tilde{x} \right] \right\|_2, \hspace{1cm} (131)$$

which is exact the same as (89). Similar results can be derived for Lemma 12. Therefore, the conclusions and proofs of Lemma 11 and Lemma 12 does not change at all.

Additionally, fixing the second-layer weights and only training the hidden layer is the state-of-the-art practice in analyzing two-layer neural networks [Arora et al., 2019b; Allen-Zhu et al., 2019; Safran & Shamir, 2018; Li & Liang, 2018; Brutzkus & Globerson, 2017; Oymak & Soltanolkotabi, 2018; Zhang et al., 2019]. Additionally, as indicated in [Safran & Shamir, 2018], training a one-hidden-layer neural network with all $v_j$ fixed as 1 has intractable many spurious local minima, which indicates that training problem is not trivial.

---

3To see this, one can view $|v_j^*| w_j^*$ as the new ground truth weights, and the goal for this paper is to recover the new ground truth weights.